RICE UNIVERSITY

Specificity in the Druggable Kinome: Molecular Basis and its Applications

by

Xi Zhang

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Ariel Fernández, Professor
Karl F. Hasselmann Chair in Engineering
Department of Bioengineering
Rice University

Michael W. Deem, John W. Cox Professor
Department of Physics and Astronomy
Department of Bioengineering
Rice University

Jianpeng Ma, Associate Professor
Bioengineering at Rice University
Biochemistry at Baylor College of Medicine

Laura Segatori, T.N. Law Assistant Professor
Chemical and Biomolecular Engineering
Rice University

HOUSTON, TEXAS

JANUARY 2009

Abstract

Specificity in the Druggable Kinome: Molecular Basis and its Applications

by

Xi Zhang

Rational design of kinase inhibitors remains a challenge partly because there is no clear delineation of the molecular features that direct the pharmacological impact towards clinically relevant targets. In this thesis, we focus on a structural marker and construct a kinase classifier that enables the accurate prediction of pharmacological differences. Our indicator is a microenvironmental descriptor that quantifies the propensity for water exclusion around preformed polar pairs. The results suggest that targeting polar dehydration patterns heralds a new generation of drugs that enable a tighter control of specificity than designs aimed at promoting ligand-kinase pairwise interactions.

As an application of the structural marker, we introduce a computational screening approach which provides a tool for extensive screening that uses experimentally obtained small-scale profiles as input data and makes predictions for a larger kinase set. These predictions result from a propagation of the reduced profile, exploiting a structural comparison of kinases based on a feature-similarity matrix. The comparison focuses on a molecular marker for specificity and promiscuity of kinase inhibitors. Our approach enables the computational high-throughput screening of entire libraries of compounds to search for suitable leads, mapping their inhibitory impact on a sizable sample of the human kinome.

Yet another application of the structural marker is advocated by illustrating its cleaning efficacy. In this regard, we reassess the possibility to turn multi-target drugs into real clinical opportunities through judicious redesign. A general cleaning strategy, which adopts the structural marker as redesigning intruction, is proposed and exemplified by a workable approach.

Acknowledgments

I thank my advisor, Dr. Ariel Fernández, for his support and insightful ideas. This thesis would not have been possible without his help. Thanks to my other thesis committee members Drs. Michael Deem, Jianpeng Ma and Laura Segatori for consenting to be on my committee. Thanks to Dr. Alejandro Crespo, Jianping Chen and Natalia Pietrosemoli for their kind help and valuable insights.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Kinase targeting is a central theme in drug discovery and molecular cancer therapy (Bain *et al.*, 2003; Druker, 2004; Hopkins *et al.*, 2006; Huse and Kuriyan, 2002; Knight and Shokat, 2005; Vieth *et al.*, 2004), but a structural basis for rational design appears to be inconclusive (Bain *et al.*, 2003; Hopkins *et al.*, 2006; Huse and Kuriyan, 2002; Vieth *et al.*, 2004). In practice, most ligands or drug leads are actually discovered through large-scale experimental screenings (Drews, 2000; Bleicher *et al.*, 2003). This is part of the reason why development of new drugs remains an expensive process. While the paradigm of target specificity (Bain *et al.*, 2003; Huse and Kuriyan, 2002) may be shifting to controlled multi-target impact (Hopkins *et al.*, 2006), the structural factors determining these possibilities are not yet fully understood, in spite of notable progress. For instance, the accessible nonpolar surface, frequently invoked to assess protein associations (Chothia, 1974; Whittle and Blundell, 1994), actually fosters promiscuity (Feng *et al.*, 2005; Hopkins *et al.*, 2006), as shown in the work presented in this thesis.

The main subject of the research presented in this thesis is to investigate the promiscuity and specificity in the druggable kinome. In this regrad, the research includes three major parts. First, we managed to identify the molecular basis of the promiscuity and specificity of the kinase inhibition. Based on our discovery, we developed two practical applications that exploit the structural feature governing the specificity of inhibition. That is, we developed an *in silico* drug profile predictor and a technique to turn promiscuous kinase inhibitors into safer drugs.

As the first step, we focus on identifying the dominant molecular feature that directs nonpromiscuous drug targeting. The identification is based on the comparison of a series of metrics that measure the differences between kinases in different senses respectively. One metric is defined to quantify the differentness between two kinases in the sense of their affinity profiles for kinase inhibitors. Several other metrics are defined to quantify the differentnesses between two kinases in their structural attributes. The essential idea to identify the molecular feature governing specific drug targeting is as follows. For any of such specificity-governing molecular features, the metric measuring the differences between kinases in the sense of this feature should be correlated with the metric quantifying the differentness between kinases in the sense of their affinity profiles for kinase inhibitors. Starting from this straightforward idea, we found a special type of hydrogen bond as the molecular basis for specificity in kinase inhibitors.

Promiscuity is operationally defined based on extensive drug screening (Fabian *et al.*, 2005) as significant cross reactivity (dissociation constant $K_d < 100nM$) extended over 30% of the sampled kinome. The problem is complicated due to the scarcity of affinity-profiled kinases with reported structure: To date, only about 20% of the human kinome

(about 100 out of the 520 discovered protein kinases) is reported in the PDB. Furthermore, kinase homology models are not useful to make sequence-based inferences since the level of sequence identity across the kinase superfamily is typically low (<30%, Manning *et al.*, 2002). Instead, we take advantage of the high degree of conservation of kinase folds, arising from their common ancestry. Thus, reliable sequence-based attributes such as disorder propensity (Braken *et al.*, 2004) are here mapped onto structurally threaded models (Fernández and Berry, 2004) to make inferences about drug specificity/promiscuity.

We search for a sequence-based attribute that enables a classification of kinase space that accurately reproduces similarities/differences in affinity profiling against a drug background (Fabian *et al.*, 2005). Our methodology could be implemented because the relevant classification of kinases was performed by comparing the regions deprived of adequate packing or intramolecular dehydration, i.e. the loopy regions, which are precisely the markers differentiating kinases at the structural level (Huse and Kuriyan, 2002).

Having identified structural marker governing specificity, we further exploited the discovery and introduced a comparison of the structural marker patterns of kinases including purported targets and experimentally confirmed targets. Based on this comparison, we developed a predictive profiler in which the comparison is adopted as surrogate for the differences in their pharmacological behavior. The core of our affinity-profile predictor involves determining a linear propagator of profiling data. This propagator consists of the structure-based estimation of pharmacological distances across kinases. Once the propagator is computed, the inference of affinity profiles for test drugs becomes a problem in distance geometry.

While the first application is adopting the structural marker in a macro manner, the

second application is conducted in a micro (case-by-case) manner. In this application, we come up with a general strategy to clean promiscuous inhibitors by introducing "wrapping" chemical modifications that preserve the drug chemotype while targeting unique dehydrons. As an operational illustration, we focus on redesigning a promiscuous compound, EKB-569, to enhance its selectivity significantly.

The thesis is organized as follows:

In chapter 2, protein structures in four levels are briefly introduced and then kinase structure is specifically described. In the following section, several relevant molecular attributes of protein are discussed: First, we formally define a region, the nonpolar hull, enabling comparison of targeted exposed nonpolar regions across different kinases. Subsequently, we investigate the dehydration propensity of protein surface. A means is introduced to calculate dehydration propensities on polar-paired regions on the protein surface. Based on the concept of dehydration propensity, the so-called solvent-accessible hydrogen bond, or dehydron, is defined. In order to expand our work to the kinases not reported in the PDB database, we introduce a sequence-based means to calculate the dehydration propensities by inferring from the disorder score (Braken *et al.*, 2004), an accurate sequence-based attribute that indicates the propensity of a chain window to be structurally disordered. At last, another type of region, the environmental hull, is defined in order to compare polar dehydration propensities across targets.

In Chapter 3, we first determine the type of molecular similarity that promotes promiscuity whenever targeted. Specifically, we demonstrate that solvent-exposed nonpolar regions engaged in ligand association foster promiscuity. Then, we examine other targeted features in molecular design aimed at inducing pair-wise interactions between ligand and

kinase and show that the high degree of conservation of the partner groups on the protein surface does not enable a cogent control of specificity. Subsequently, we demonstrate that the dehydration propensities is responsible for controlling specificity. To carry out the analysis at a sequence level, we adopt the environmental alignment technique introduced in Chapter 2, which enables the identification of residues whose microenvironment are likely to be perturbed by ligand association. Such residues are identified by aligning the kinase sequence against a background of sequences of homologous kinase-ligand complexes reported in the PDB.

In Chapter 4, we discuss the first application of the structural marker governing specificity, i.e., the *in silico* drug profile predictor. To construct this drug profile predictor, we first show how the profile prediction problem is re-cast in linear algebra terms. We then solve the corresponding linear algebra problem. To carry out the prediction in practical cases, some optimizing steps are needed, which is addressed in this chapter. To validate the predictor, we test the predictions versus published experimental data with testing results presented in this thesis. Also, we assess the performance of our predictor by comparing it with other computational predictors.

Chapter 5 presents another application of the structural marker, that is, the strategy to clean promiscuous inhibitors by introducing "wrapping" chemical modifications that preserve the drug chemotype, while targeting unique dehydrons. First, we discuss the possibility of cleaning cross-reactive drugs by exploiting the structural marker as selectivity filters. Then we analyze the molecular basis of imatinib-redesigning to curb its potential cardiotoxicity and further propose the general strategy of cleaning promiscuous inhibitors. To illustrate the operational value of this approach, we focus on redesigning a promiscuous

drug, EKB-569, to enhance its selectivety significantly. In details, we describe how to trace the specific molecular basis for EKB-569 promiscuity, the modifications to remove such sources of promiscuity and the modification to promote selectivity through targeting a non-conserved dehydron.

Chapter 2

Protein Structure and Relevant Molecular Attributes

2.1 Protein Structure

Proteins are large organic compounds made of amino acids arranged in a linear chain and joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. They function as catalysts, transporters and store places of other molecules such as oxygen, mechanical supporter, immune protector, and growth and differentiation controller. Protein structure can be described at four levels: The primary structure refers to the amino acid sequence. The secondary structure refers to the conformation adopted by local regions of the polypeptide chain. Tertiary structure describes the overall folding of the polypeptide chain. Finally, quaternary structure refers to the specific association of multiple polypeptide chains to form multisubunit complexes. In this section we

only briefly describe the protein structures. More detailed information of this is available in any textbook of biochemistry, e.g. Berg *et al.* (2002); Nelson and Cox (2005).

2.1.1 Amino acid and nonpolar group

Proteins are linear polymers of amino acids, each of which consists of a central tetrahedral carbon atom linked to an amino group, a carboxylic acid group, a distinctive side chain, and a hydrogen atom (figure 2.1). All natural proteins are constructed from the same set of 20 amino acids (table 2.1). The side chains of these 20 building blocks vary tremendously in size, shape, the presence of functional groups and polarity. According to side chain polarity, amino acids can be categorized as *polar* or *nonpolar* (cf. table 2.2). Generally, polar side chains are hydrophilic while nonpolar side chains are hydrophobic.

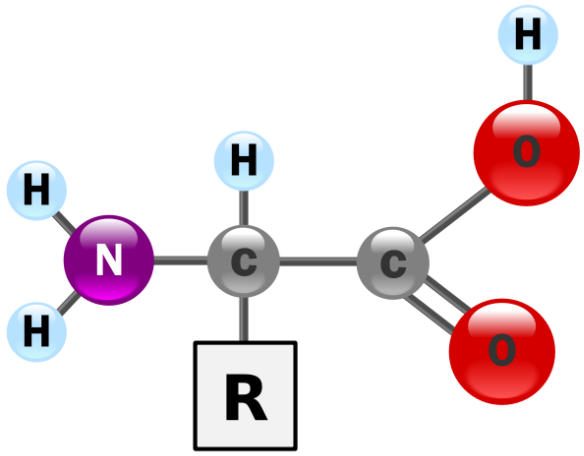
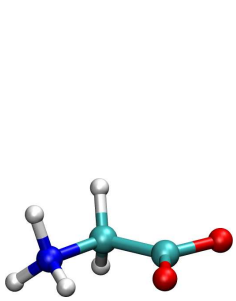
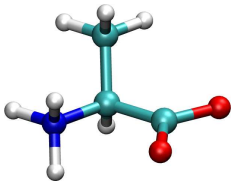
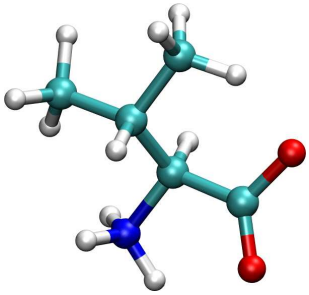


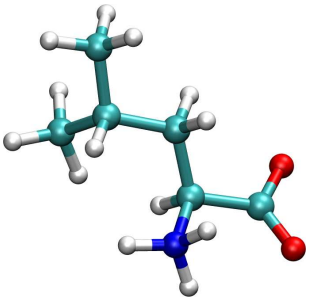
Figure 2.1: The general structure of an amino acid, with the central α carbon atom in the middle, the amino group on the left, the carboxylic group on the right, the side chain on the bottom (R) and the hydrogen atom on the top.

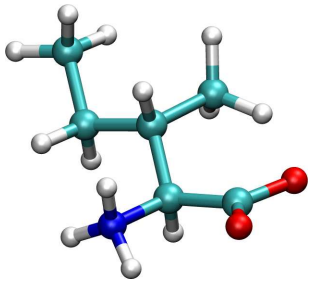
Table 2.1: The 20 amino acids

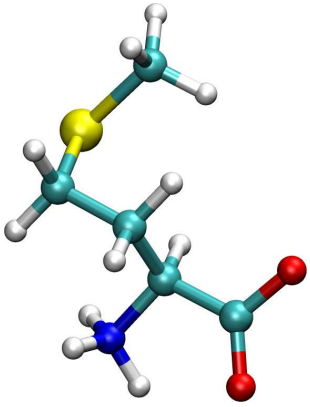

Glycine

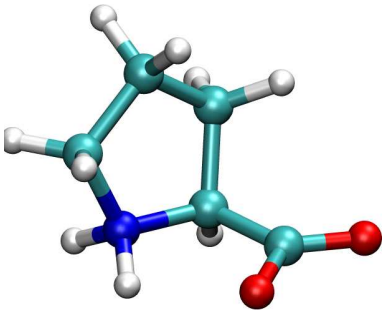

Alanine

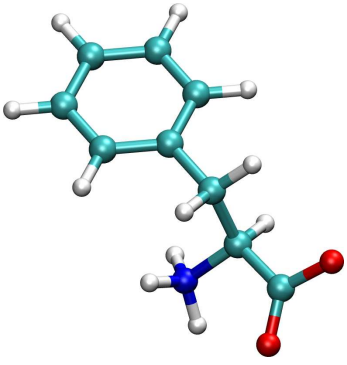

Valine

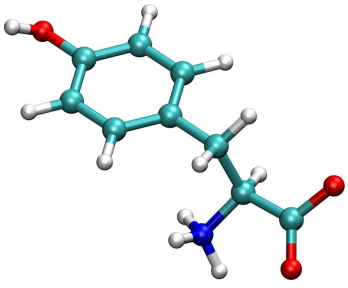

Leucine

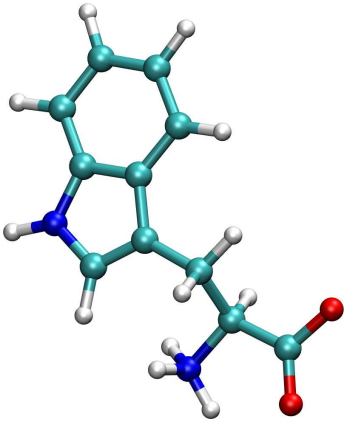

Isoleucine

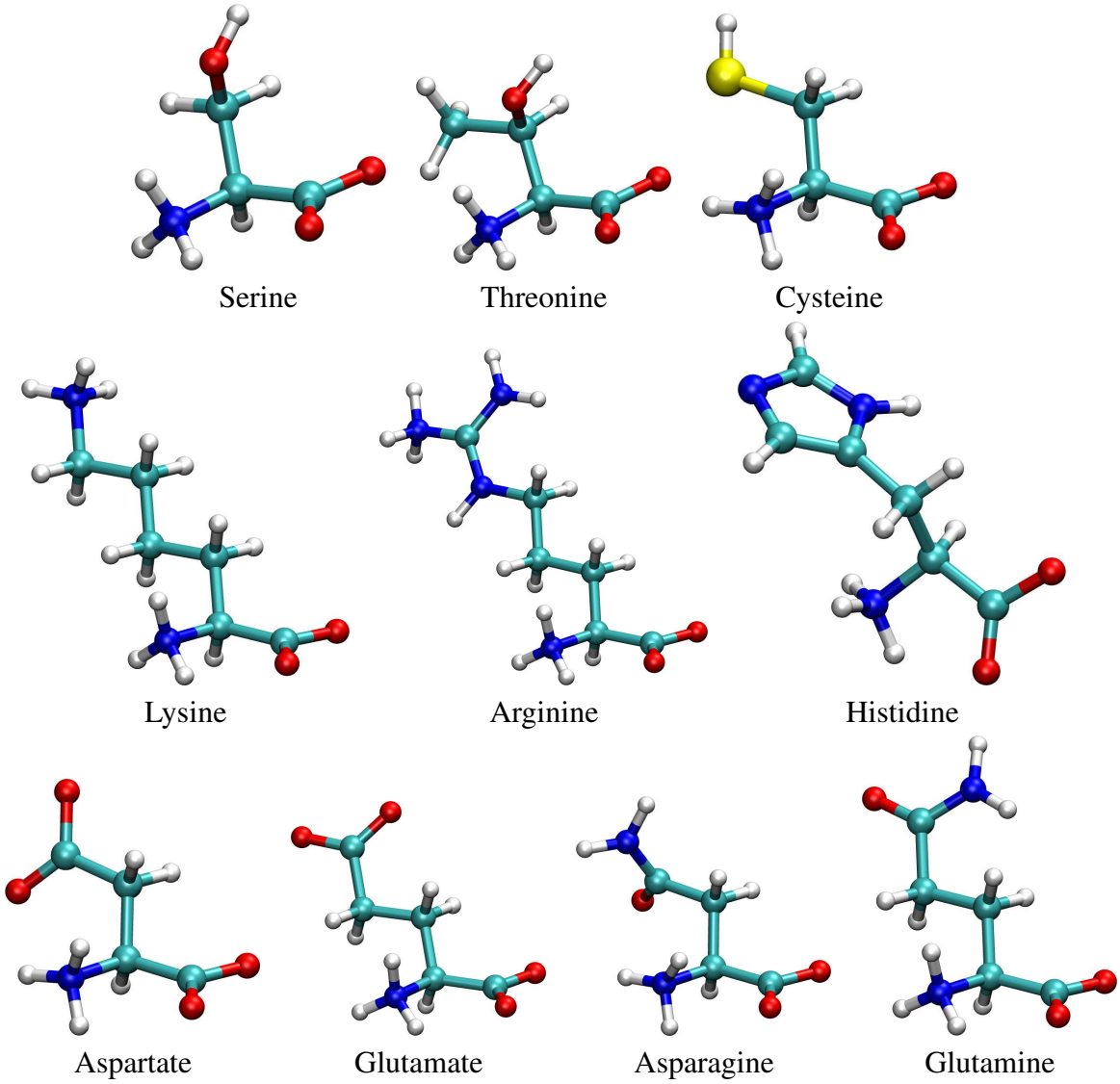

Methionine


Proline


Phenylalanine


Tyrosine


Tryptophan



One of the essential features of amino acids is the nonpolar groups within the molecules. Nonpolar groups are defined as the carbonaceous groups that are within protein side chains and are not directly connected to any polar groups (O, N, S). The nonpolar groups of all the 20 amino acids are listed in table 2.2. In this table, each nonpolar group is represented

by the C atom within the group, which is named following the naming rules used in PDB database[1]. Note that some polar amino acids, e.g. lysine and arginine, do have nonpolar groups within their side chains. On the other hand, glycine, which is nonpolar, does not have any nonpolar group since its side chain is too small, only a hydrogen atom. However, generally a nonpolar amino acid has more nonpolar groups than a polar amino acid does, which is intuitive. The nonpolar groups constitute the atomic basis of our work because they are able to wrap the hydrogen bonds between main chain atoms and thus protect the wrapped hydrogen bonds from hydration. Notice that not only amino acids have nonpolar groups, but ligands could also have them, that is, the carbonaceous groups within ligand molecules not directly connected to any polar groups. Thus, ligands, e.g. drug molecules, may also be able to wrap hydrogen bonds between main chain atoms when they properly bind to proteins.

2.1.2 Protein structures in four levels

Amino acids are linked by peptide bonds to form polypeptide chains. The linkage is formed by amide bonds between the carboxyl group of one amino acid and the amino group of the next. This linkage, i.e. the so-called peptide bond, has several important properties. First, it is resistant to hydrolysis so that proteins are remarkably stable kinetically. Second, the peptide group is planar because the C-N bond has considerable double-bond character. Third, each peptide bond has both a hydrogen bond donor (the NH group) and a hydrogen bond acceptor (the CO group). Hydrogen bonding between these backbone groups is a distinctive feature of protein structure. Lastly, the peptide bond is uncharged, which allows

[1] http://www.wwpdb.org/docs.html

Table 2.2: Side chain polarities and nonpolar groups of all the 20 amino acids[†]

Amino Acid	Side chain polarity	nonpolar group
Glycine	nonpolar	
Alanine	nonpolar	CB
Valine	nonpolar	CB, CG1, CG2
Leucine	nonpolar	CB, CG, CD1, CD2
IsoLeucine	nonpolar	CB, CG1, CG2, CD1
Methionine	nonpolar	CB
Proline	nonpolar	CB, CG
Phenylalanine	nonpolar	CB, CG, CD1, CD2, CE1, CE2, CZ
Tyrosine	nonpolar	CB, CG, CD1, CD2, CE1, CE2
Tryptophan	nonpolar	CB, CG, CD2, CE3, CZ3, CZ2, CH2
Serine	polar	
Threonine	polar	CG2
Cysteine	polar	
Asparagine	polar	CB
Glutamine	polar	CB, CG
Aspartate	polar	CB
Glutamate	polar	CB, CG
Lysine	polar	CB, CG, CD
Arginine	polar	CB, CG
Histidine	polar	CB

[†] Nonpolar groups are represented by the C atoms within them,which are named following the naming rules of PDB: CA (A representing α) is α-C in the main chain; CB (B representing β) is β-C which is the C atom closest to CA; CG is γ-C and so on. When there are two or more symmetric C atoms, they are denoted by numbers, e.g. CD1 and CD2 are two δ-C's. The convention using greak letters to identify carbons is described in page 76 of Nelson and Cox (2005).

proteins to form tightly packed globular structures having significant amounts of the backbone buried within the protein interior. Because they are linear polymers, proteins can be described as sequences of amino acids. Such sequences are written from the amino to the carboxyl terminus.

Once the amino acids are linked into polypeptide chains, the chains can then fold into

secondary structures, such as α helices, β sheets, and turns and loops. α-helix and β sheet are the two major elements of protein's secondary structure. In an α helix, the polypeptide chain twists into a tightly packed rod, within which the CO group of each amino acid is hydrogen bonded to the NH group of the amino acid four residues along the polypeptide chain. In a β strand, the polypeptide chain is nearly fully extended. Two or more β strands connected by NH-to-CO hydrogen bonds come together to form β sheets. Turns and loops are usually the parts between α helices or β sheets and are generally more flexible than the latter two.

Based on the secondary structures, proteins form their tertiary structures by folding into compact globular shapes. The tertiary structures of water-soluble proteins have features as follows: a) an interior formed of amino acid with hydrophobic side chains and b) a surface formed largely of hydrophilic amino acids that interact with the aqueous environment. The driving force for the formation of the tertiary structure of water-soluble proteins is the hydrophobic interactions between the interior residues.

Furthermore, polypeptide chains can assemble into multisubunit structures, i.e. quaternary structures. Quaternary structure can be as simple as two identical subunits (i.e. dimer) or as complex as dozen of different subunits. In most cases, the subunits are held together by noncovalent bonds.

2.1.3 Protein kinase structure

In this thesis, we focus on one special type of protein: protein kinase. Protein kinases are quintessential signal transducers, and thus their inhibition becomes a central strategy to block specific signaling pathways, as often needed for therapeutic reasons (Taylor and

Radzio-Andzelm, 1997; Donato and Talpaz, 2000; Dancey and Sausville, 2003; Fabbro and Garcia-Echeverria, 2002; Gabriele and King, 2001; Myers *et al.*, 1997). In fact, a number of diseases, including cancer, diabetes, and inflammation, are linked to perturbation of protein kinase-mediated cell signaling pathways (Noble *et al.*, 2004). The human genome encodes 518 protein kinases (Manning *et al.*, 2002) that share a catalytic domain conserved in structure but which are notably different in how their catalysis is regulated. The catalytic domain of protein kinase is the main focus of this thesis. One typical catalytic domain of a protein kinase (figure 2.2) consists of two lobes (N-terminal lobe and C-terminal lobe) and the hinge part. The N-terminal lobe consists of a β sheet and one conserved α helix (helix C). The C-terminal lobe is largely helical. The hinge region connects two lobes, through the so-called catalytic loop or C-loop. C-loop, together with the activation loop (A-loop) from the C-terminal lobe and the phosphorylation loop (P-loop) from the N-terminal lobe, forms the ATP-binding site, which is also the binding site of ligand.

From a structural perspective, the kinase ATP pockets (Huse and Kuriyan, 2002; Noble *et al.*, 2004) are obvious targetable features. However, due to the common evolutionary ancestry of kinases, ATP-binding sites are highly conserved in shape and amino acid composition across the superfamily (see the illustrative example in figure 2.3), making it hard to achieve selectivity or control the inhibitory impact (Hopkins *et al.*, 2006; Feng *et al.*, 2005). Thus, selectivity becomes a major challenge in kinase-inhibitor development. And this is exactly the problem dealt with in this thesis.

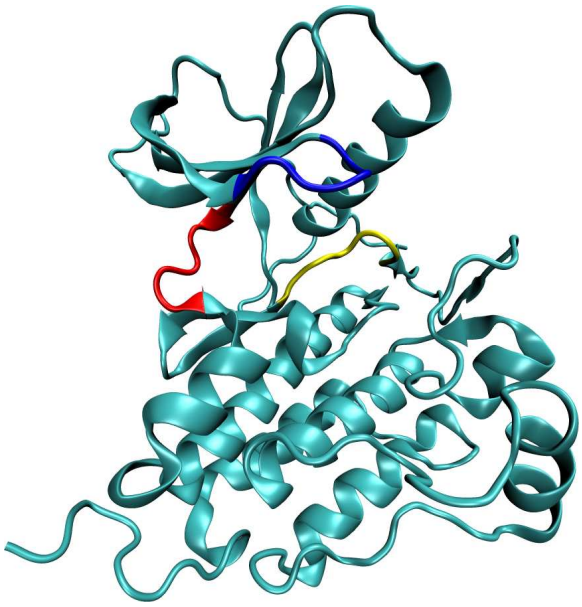


Figure 2.2: The structure of the catalytic domain of EGFR (1M17.pdb rendered by VMD). It consists of two lobes (N-terminal lobe on the top and C-terminal lobe on the bottom) and the hinge region (in red color). The N-terminal lobe consists of a β sheet and one conserved α helix (helix C). The C-terminal lobe is largely helical. The hinge region connects two lobes, through the so-called catalytic loop or C-loop (shown in red). C-loop, together with the activation loop (or A-loop, indicated in yellow) from the C-terminal lobe and the phosphorylation loop (or P-loop, indicated in blue) from the N-terminal lobe, forms the ATP-binding site, which is also ligand-binding site.

2.2 Molecular attributes of protein

In this section we introduce a series of molecular attributes of protein in preparation for the discussion in the next chapter.

2.2.1 Nonpolar hull

One important feature of protein is its polar/nonpolar pattern, i.e. the spatial distribution of polar amino acids and nonpolar amino acids within the protein. In order to assess

Figure 2.3: Structural alignment of FAK (focal adhesion kinase, pdb code 2ETM), a major cancer drug target, and INSR (insulin receptor kinase, pdb code 1GAG), a target to be avoided at any cost in molecular therapy. The two structures are aligned by DaliLite (http://www.ebi.ac.uk/DaliLite/) and rendered with VMD (http://www.ks.uiuc.edu/Research/vmd/). Only backbones are indicated for clarity. The structure similarity (RMSD$\sim$ 0.9 Å) between the two kinases may lead to life-threatening cross reactivity since INSR is indispensable kinase mediating the metabolic functions of insulin.

differences in the nonpolar patterns of the exposed regions of kinase targets that interact with different ligands, we first define a common region, named the nonpolar hull. A residue a is defined as making contact with a ligand L within a PDB-reported complex if a side-chain heavy atom (H excluded) is found to be within 3.6 Å (upper bound for any bond length) of a heavy atom in the ligand. The nonpolar hull for protein chain i, $H_{np}(i)$, is dependent on a 'structural background set' of chains, $S(i)$, which includes all homolog chains that align with chain i (Higgins *et al.*, 1996) for which there are protein-ligand complexes

with reported structure. Thus, the nonpolar hull is defined as

$$H_{np}(i) = \cup_{j \in S(i)} \Phi_i(R_{np}(j)), \tag{2.1}$$

where Φ is the alignment operator such that $\Phi_i(a)$, with $a \in$ chain j, is the residue in chain i that aligns with residue a in chain j, and $R_{np}(j)$ is the set of nonpolar residues (cf. table 2.2) in chain j in contact with its respective ligand L_j. For any pair i, j, the following property holds:

$$\Phi_i(H_{np}(j)) = H_{np}(i), \tag{2.2}$$

which enables a comparison of kinases by examining differences in nonpolar hulls. Figure 2.4 illustrates the definition of nonpolar hull and 2.5 shows an example: the nonpolar hull of the active fold of pregnane X receptor (PXR).

2.2.2 Hydration environment and dehydration propensity of hydrogen bond

Now we introduce another molecular attribute of protein, i.e., the hydration microenvironment of hydrogen bonds in protein. A preformed hydrogen bond microenvironment may be determined from the atomic coordinates of the protein by calculating the extent of intramolecular dehydration, ρ, quantified as the number of the side chain carbonaceous nonpolar groups (cf. table 2.2) within a dehydration domain (cf. figure 2.6). This domain consists of two intersecting balls of radius 6.0 Å ($\sim$ width of three solvation layers (Fernández and Berry, 2004)) centered at the α-carbons of the hydrogen-bond paired residues.

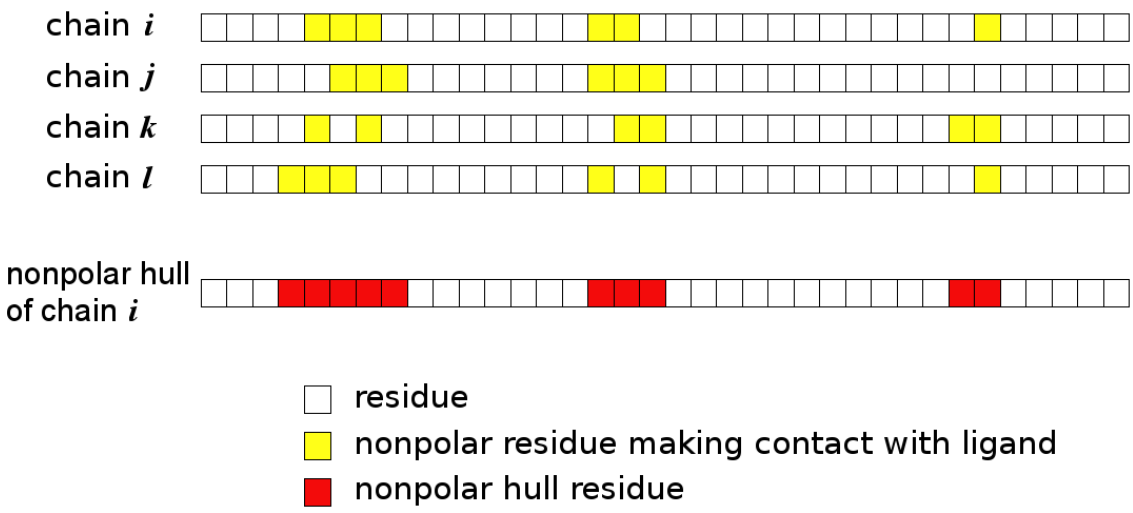


Figure 2.4: Nonpolar hull. Each grid represents a redidue and each grid chain represents a protein chain. The chains are aligned with each others. The yellow grids are the nonpolar residues and the nonpolar hull is colored in red.

In soluble protein domains, at least two thirds of the backbone hydrogen bonds lie in the range $\rho = 26.6 \pm 7.5$. A subsequent concept based on the dehydration propensity is solvent-accessible hydrogen bond (SAHB), or sometimes named as *dehydron*. In the following part of this thesis, both of the two names will be used. SAHB is defined such that the extent of intramolecular dehydration (ρ) of a solvent-accessible hydrogen bond lies in the tails of the distribution, i.e., with 19 or fewer nonpolar groups in its microenvironment. That is, its ρ-value is below the mean, minus one Gaussian dispersion. Such bonds constitute dehydration-propensity hot spots as demonstrated before (Fernández, 2004; Fernández and Berry, 2004).

The SAHBs may be determined directly from a PDB file using the program YAPView[2].

[2]YAPView is developed in a CS lab of University of Chicago and can be downloaded from website:

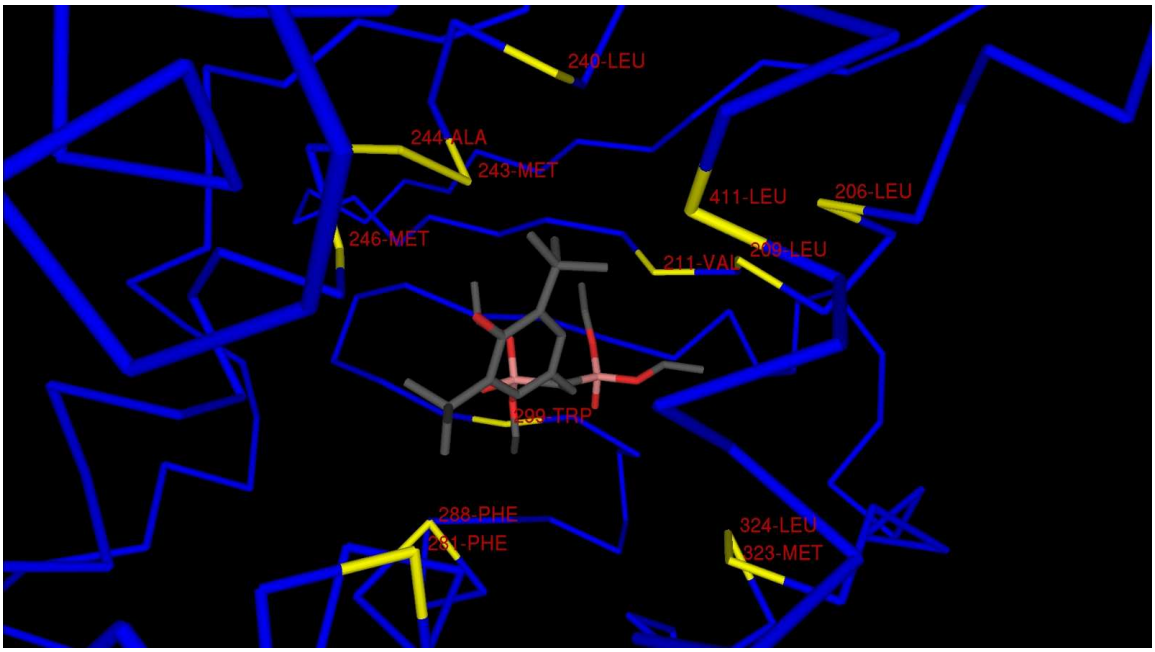


Figure 2.5: The nonpolar hull of the active fold of pregnane X receptor (PXR) in complex with SR12813 (PDB.1ILH). The virtual bonds between α-carbons are depicted in blue, while the residues in the hull are shown in yellow.

This program is inspired in earlier desolvation calculations (Fernández *et al.*, 2002). Within YAPview, the SAHBs are identified by loading the PDB file, choosing a structure display/representation and enabling a desolvation calculation. The latter is needed to determine the extent of intramolecular dehydration of hydrogen bonds. This operation requires the selection: Configuration → General Options → Desolvation. Thereafter, one needs to enable the desolvation calculator and select the appropriate parameters, especially desolvation radius and desolvation threshold, according to the indications given above.

Thus, YAPView displays the dehydration-propensity hot spots directly on the protein surface: The hydrogen bonds that are poorly dehydrated intramolecularly, that is, below

http://sourceforge.net/project/showfiles.php?group_id=133896&package_id=225002

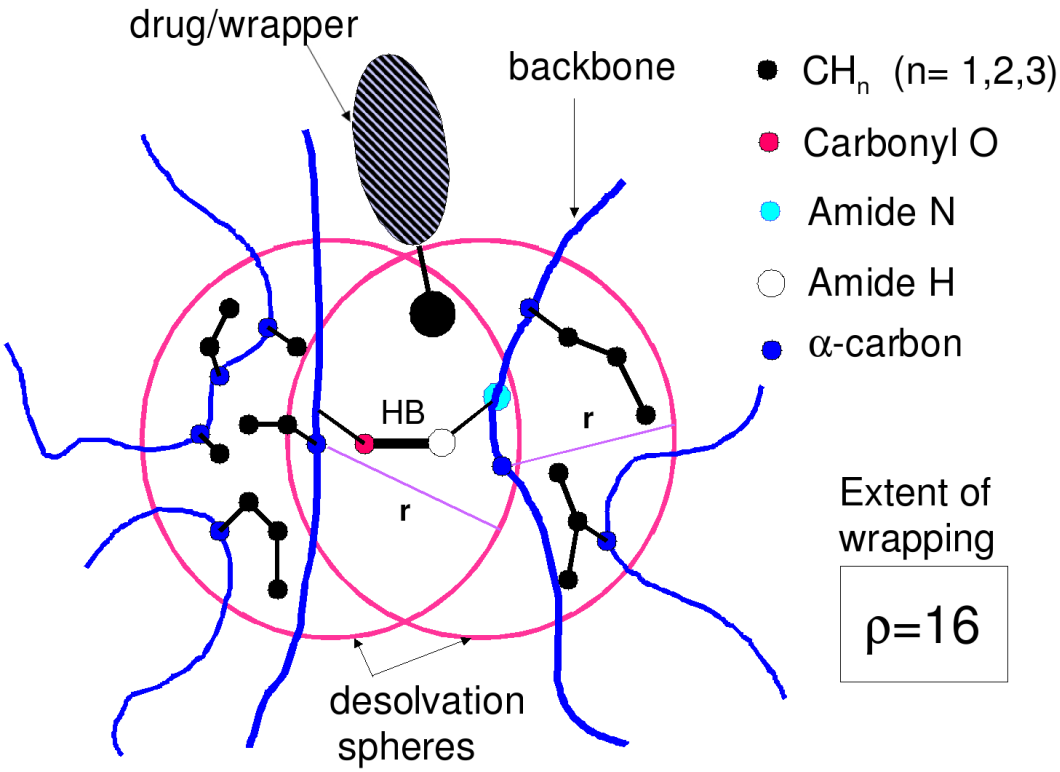


Figure 2.6: Hydrogen-bond microenvironment. Intramolecular dehydration, ρ, is quantified as the number (16 in this case) of side-chain nonpolar groups (black disks) within the dehydration domain (the two intersecting spheres) defining the microenvironment of a particular hydrogen bond.

the pre-selected threshold are shown in green in the structure display and promote the exclusion of surrounding water.

2.2.3 Sequence-based prediction of dehydration propensity

While YAPview can directly calculate the hydration propensities of hydrogen bonds and thus identify SAHBs from PDB file, our analysis is not constrained to PDB-reported kinases. Robetta/Rosetta predictions of active structures (Bonneau *et al.*, 2002; Chivian *et al.*, 2005) become reliable given the extent of PDB representation (27%) of paralogs within the kinase superfamily and since SAHBs may be directly inferred from sequence (Fernández and Berry, 2004), and contrasted with the structure predictions for mutual validation. Such SAHB inferences make use of a strong correlation between the extent of dehydration of the preformed hydrogen bonds and the disorder score (Braken *et al.*, 2004), an accurate sequence-based attribute that indicates the propensity of a chain window to be structurally disordered. The correlation is maintained irrespectively of whether the structure is a Rosetta prediction or PDB reported and implies that native disorder arises essentially from the impossibility to dehydrate intramolecular hydrogen bonds. Figure 2.7 illustrates the strong correlation.

The disorder propensity is given by a score determined by the program PONDR®(Braken *et al.*, 2004), a neural-network predictor of native disorder. Only 0.4% of more than 900 non-homologous PDB proteins give false positive predictions in regions with 40 or more consecutive sites of predicted disorder. Even this 0.4% of false positives is an overestimation, as many disordered regions in monomeric chains become ordered upon ligand binding or in crystal contacts (Braken *et al.*, 2004). The false negatives error rate (11% for regions of 40 or more consecutive predicted ordered residues) is also compelling in regards to the predictor quality.

The correlation between solvent exposure of hydrogen bonds and disorder propensity

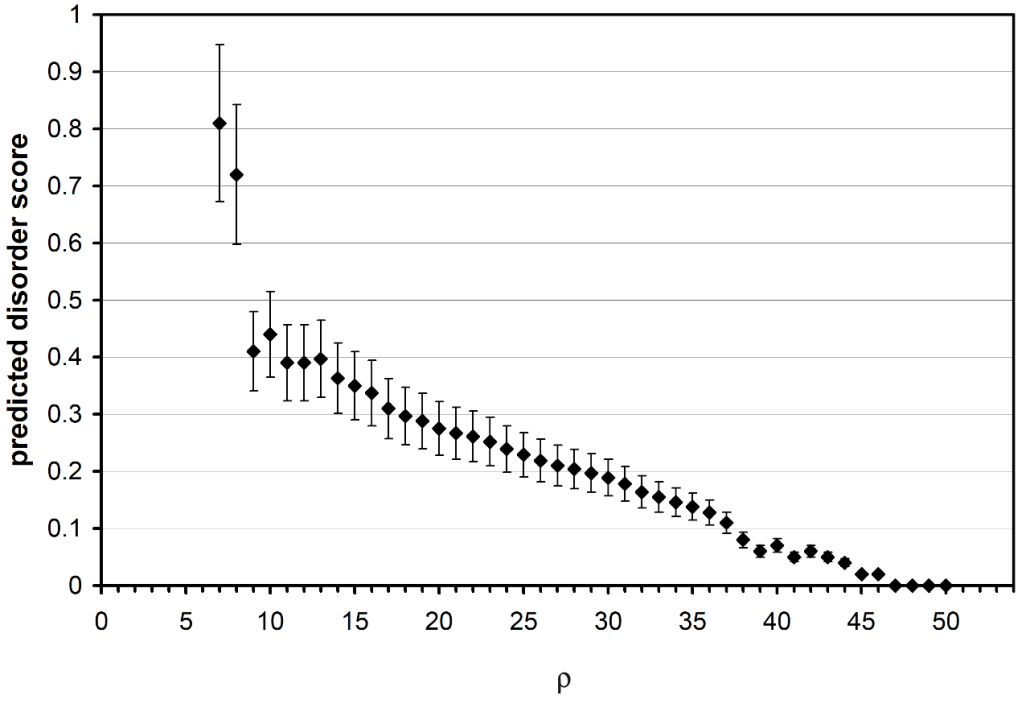


Figure 2.7: Correlation between the disorder score of a residue and the extent of intramolecular dehydration (ρ) of the backbone hydrogen bond engaging that particular residue. The disorder score on each individual residue was obtained for 2806 nonredundant nonhomologous PDB domains. Residues have been independently grouped in 46 bins of 400 residues each, according to the extent of wrapping ($7 \leq \rho \leq 52$). The average score has been determined for each bin, and the error bars represent the dispersion of disorder scores within each bin. The strong correlation between the disorder score and extent of wrapping and the dispersions obtained imply that dehydrons can be safely inferred in regions where the disorder score is above 0.35.

implies that it is possible to predict SAHBs directly from sequence (Fernández and Berry, 2004): It suffices to determine the PONDR-generated pattern associated with the desired feature. The correlation implies that the propensity to adopt a natively disordered state becomes pronounced for proteins which, due to a chain composition reflecting high hy-

drophilicity, cannot protect even minimally the backbone hydrogen bonds. Thus, we can infer the existence of SAHBs from the PONDR score (λ_d) with 92% accuracy in regions with λ_d >0.35 provided such regions are flanked by well-protected regions ($\lambda_d < 0.35$), to ensure the existence of structure. The accuracy of this sequence-based SAHB predictor was established by inferring the location of SAHBs in proteins with reported structure, for which the microenvironment of each hydrogen bond can be determined unambiguously (Fernández and Berry, 2004). The false negatives constitute 368 of the 8,215 SAHBs in a PDB database of 1,466 proteins free from structural redundancy and less than 25% sequence identity in pairwise alignment. The false positives correspond to 2721 of the 133,623 backbone hydrogen bonds examined.

2.2.4 Environmental hull and environmental alignment technique

To assess differences in the dehydration propensities of polar regions for purported kinase targets, we introduce a common region named environmental hull, which is defined in a way similar to that of nonpolar hull (section 2.2.1). First, the set $R_{env}(j)$ is defined for protein chain j as the set of residues paired by SAHBs in chain j within a protein-ligand complex with reported structure subject to the following condition: ligand L_j contributes to the dehydration of the SAHB, that is, it has some carbonaceous nonpolar groups within the dehydration domain of the SAHB. Then, the set $E_{env}(j)$ is defined as the set of residues from chain j that contribute to the dehydration microenvironment of a SAHB contained in $R_{env}(j)$, that is, they are either paired by the SAHB or they contain a side-chain nonpolar group within the microenvironment of the SAHB. Then, the environmental hull for chain j, $H_{env}(j)$, is defined as the union of the residues in chain j that align with residues framing

the environments of SAHBs that in turn are environmentally affected by ligands in PDB-reported complexes:

$$H_{env}(j) = \cup_{i \in S(j)} \Phi_j(E_{env}(i)) \tag{2.3}$$

where notation has been followed consistently and the structural background S(j) is defined above. As with nonpolar hulls, for any pair i, j, the following property also holds:

$$\Phi_i(H_{env}(j)) = H_{env}(i). \tag{2.4}$$

As a property similar to that of nonpolar hull, the equality is needed to actually compare environments of different proteins. Figure 2.8 illustrates the definition of enrironmental hull and figure 2.9 shows a real example of this definition.

2.3 Summary

The main purpose of this chapter is to introduce the background knowledge for the research presented in this thesis. We first briefly introduced protein structures in four levels, and specifically, the kinase domain of kinase proteins, which is the central object of our research.

In preparation for the following investigations, several relevant molecular attributes of protein have been defined and described. First, we formally defined a region, the nonpolar hull, enabling comparison of targeted exposed nonpolar regions across different kinases. Subsequently, we investigate the dehydration propensity of protein surface. A means is introduced to calculate dehydration propensities on polar-paired regions on the protein

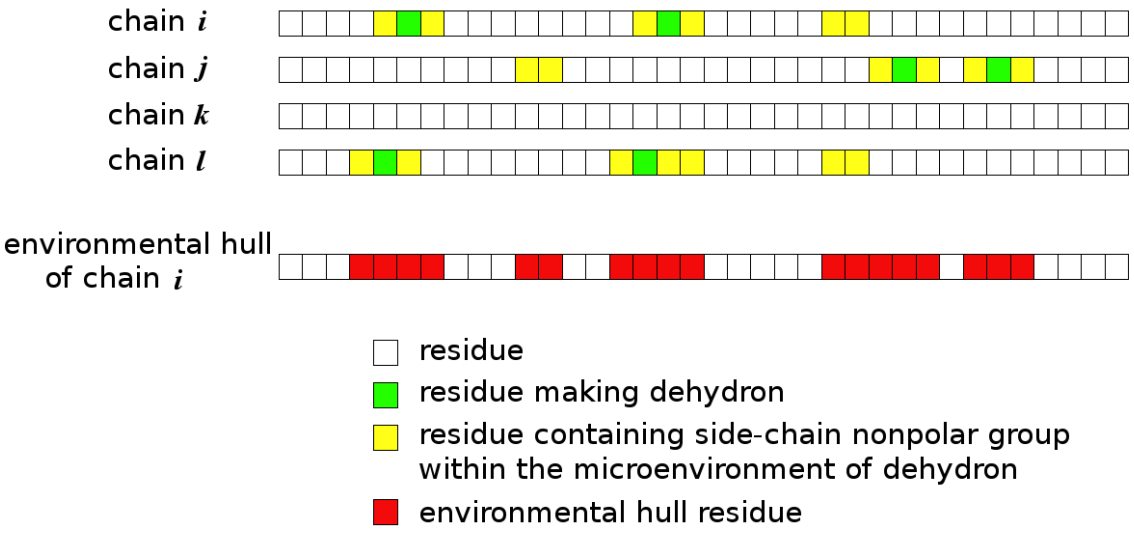


Figure 2.8: Environmental hull. Each grid represents a redidue and each grid chain represents a protein chain. The chains are aligned with each others. The green grids represent the residues forming dehydons (SAHBs) while the yellow grids represent the residues containing side-chain nonpolar group within the microenvironment of any dehydron. The environmental hull is colored in red.

surface. Based on the concept of dehydration propensity, the so-called solvent-accessible hydrogen bond, or dehydron, is defined. This is the most important molecular basis for the researches presented in the following chapters.In order to expand our work to the kinases not reported in the PDB database, we introduce a sequence-based means to calculate the dehydration propensities by inferring from the disorder score (Braken *et al.*, 2004), an accurate sequence-based attribute that indicates the propensity of a chain window to be structurally disordered. At last, another type of region within protein, the environmental hull, is defined in order to compare polar dehydration propensities across targets.

All the following chapters are based on the background knowledge introduced in this chapter. In Chapter 3, we investigate the relationships between the molecular attributes of

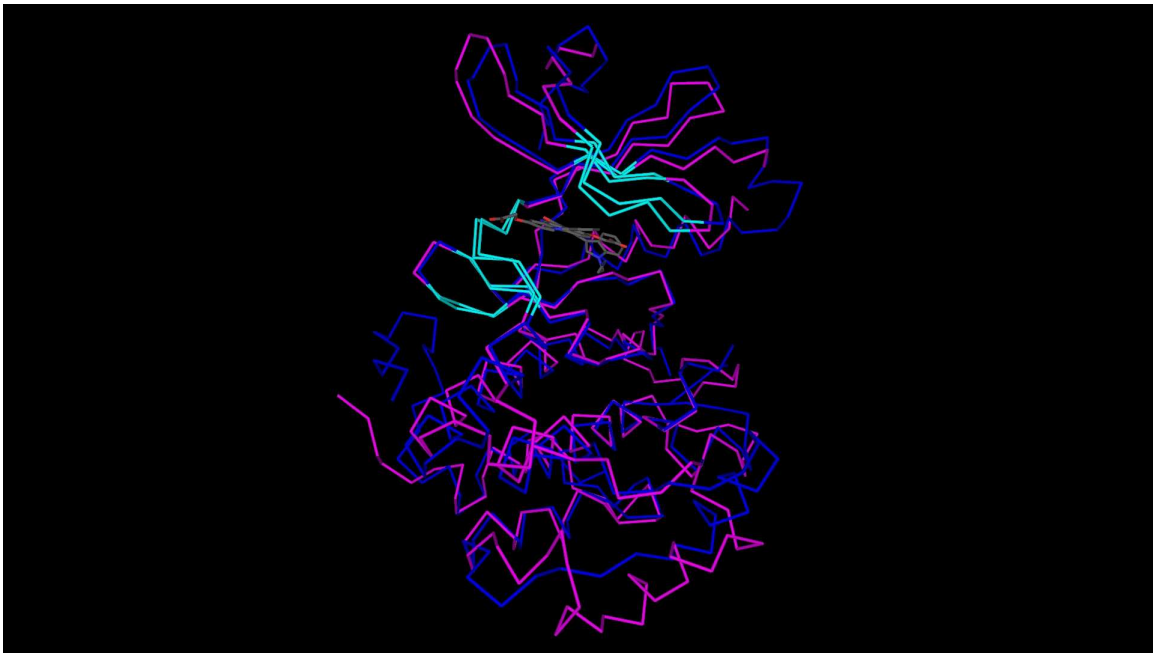


Figure 2.9: Aligned backbones for PDK1 (blue) in complex with BIM8 (PDB.1UVR) and CHK1 (lilac) in complex with 3A3 (PDB.2CGU), with the environmental hulls depicted in light blue.

kinase proteins and drugs' affinity profiles against kinase proteins and identify the molecular basis for the specificity of kinase inhibitor. In Chapter 4, the dehydron concept is further exploited to infer the pharmacological differences between kinases and thus to predict affinity profiles of drug. In Chapter 5, the dehydron is adopted as a structural marker for the manipulation of therapeutic impact in drug design/redesign.

Chapter 3

Molecular Basis for Promiscuity and Specificity

The elucidation of the molecular factors governing promiscuity and specificity in molecularly targeted drug therapy requires that we attempt to correlate different molecular attributes with available screening data for a sizable set of kinase targets. In this chapter we first reveal the molecular basis for promiscuity: pair-wise interactions between ligand and kinase target corresponding to the high conservation of the partner groups on or around the ATP-binding site of the kinase. Subsequently, we show that the SAHB defined in Chapter 2 may be targeted to promote specificity. The large assayed set adopted (Fabian *et al.*, 2005) is highly underreported in the PDB and consequently, a reliable sequence-based predictor (see section 2.2.3) of the relevant molecular attributes had to be implemented. We adopted Robetta/Rosetta predictor (Bonneau *et al.*, 2002; Chivian *et al.*, 2005) to accomplish the sequence-based predictions wherever needed and validated the results with

disorder propensity (Braken *et al.*, 2004). The predictor is based on alignments against sequences of PDB-reported kinases that are used to define the windows for comparison.

3.1 Molecular basis for promiscuity

3.1.1 Pharmacological distance

In order to elucidate the relationship between compounds' molecular attributes and their affinity profile for the kinases, we first define a pharmacological distance, d_{phar}, that quantifies differences in the affinity profiling of kinases against a background of available drugs (Fabian *et al.*, 2005). The idea behind this metric is that any factor governing the specificity of kinase inhibitor should be able to construct a quantity that is correlated with pharmacological distance. This metric is the Euclidean distance between affinity vectors with entries given in negative logarithm of dissociation constant or $\Delta G/RT$ units (ΔG = free energy change for protein-ligand association, R = universal gas constant, T = absolute temperature). The cutoff value for "no hit" affinities is $\Delta G/RT = ln10 \approx 2.3$ (i.e. $K_{d,cutoff} = 10\mu M$, Fabian *et al.*, 2005). And every "no hit" entry (with $K_d > K_{d,cutoff}$)in the screening table of (Fabian *et al.*, 2005) is assigned the value $-23.026\Delta G/RT$-unit, corresponding to a large dissociation constant $K_d = 10^{10}\mu M$. The distance is given by

$$d_{phar}(i,j) = \left[\sum_{m \in \text{drug background}} (K(i,m) - K(j,m))^2 \right]^{1/2}, \tag{3.1}$$

where $K(i,m)$, $K(j,m)$ represent respectively the negative logarithm of equilibrium constants for complexation of kinase i and kinase j with drug inhibitor m belonging to the

drug background. Figure 3.1 displays the matrix $\mathbf{D}_{phar} = [d_{phar}(i, j)]$ for all pairs (i, j) from the 119 assayed kinases. Theoretically, the *drug background* should cover all the existent drugs and thus the corresponding affinity vector space should have almost infinite dimensions. However, this is not practically feasible because of two major reasons: 1) An infinite-dimensional space cannot be handled numerically; 2) For some of the drugs, we have no profile information at all and thus the corresponding dimensions are "invisible". Practically, we only adopted a number of drug compounds to constitute the *drug backgroun*, which is in fact a subspace of the complete *drug background*. The affinity profiling adopted included 19 of the 20 drugs originally screened (Fabian *et al.*, 2005): only the promiscuous ligand staurosporine was initially excluded since it does not belong to the pharmacology realm.

3.1.2 Nonpolar pattern and promiscuity

To determine whether pharmacological differences are dictated by differences in nonpolar accessible surfaces of the targets within ligand-binding sites, a nonpolar distance, $d_{np}(i, j)$, between the affinity-assayed kinases i, j is introduced. The $d_{np}(i, j)$ is determined by differences in accessible nonpolar surface areas of the respective nonpolar hulls, $H_{np}(i)$, $H_{np}(j)$. The nonpolar hull of a kinase (section 2.2.1) is comprised of the nonpolar residues of the kinase in contact with its drug ligands, whenever the complexes are reported in PDB, and residues in the kinase (not necessarily nonpolar) that align with nonpolar residues in contact with ligands in PDB-reported ligand-kinase complexes (Figure 3.2). Introducing the hull becomes necessary to compare kinases at this level. Thus, the nonpolar distance is

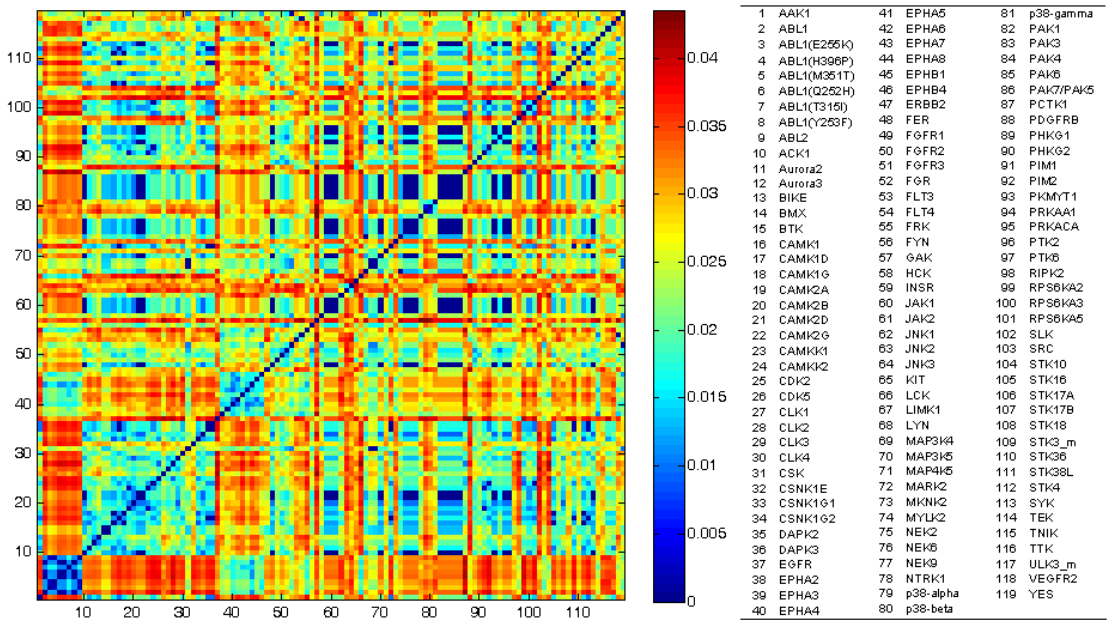


Figure 3.1: Pharmacological distance matrix $\mathbf{D}_{phar} = [d_{phar}(i,j)]$ for all pairs (i,j) from the 119 kinases assayed through affinity profiling against a background of 19 drugs (Fabian *et al.*, 2005):SB202190; SB203580; sp600125; imatinib (Gleevec); VX-745; BIRB 796; BAY-43-9006; GW-2016; gefitinib; erlotinib; CI-1033; EKB-569; ZD-6474; Vatalanib; SU11248; MLN-518; LY-333531; roscovitine/CYC202 and flavopiridol.

expressed as

$$d_{np}(i,j) = \frac{1}{[A(H_{np}(i))]} \sum_{a \in H_{np}(i)} |P(a) - P(\Phi_j(a))|, \tag{3.2}$$

where

A = nonpolar accessible area ((Fraczkiewicz and Braun, 1998);

a = generic residue in chain i;

$P(a) = 0$ if a is polar, 1 if a is nonpolar;

$\Phi_j(a)$ = residue in chain j that aligns with a in i.

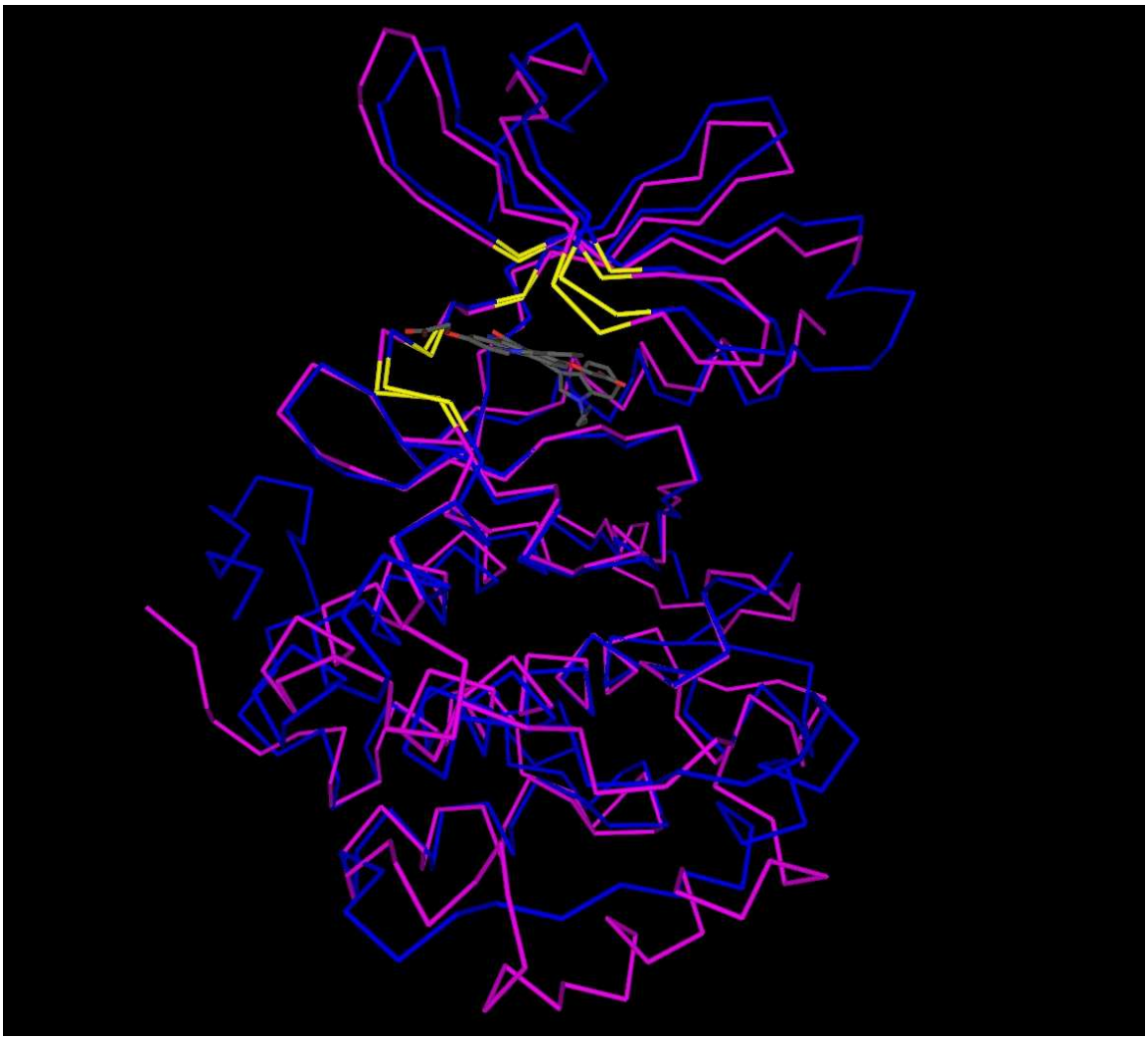


Figure 3.2: Aligned backbones(Hogue, 1997) (RMSD$\approx$ 0.33 Å) for paralog kinases PDK1 (blue) and CHK1 (lilac) in their active folds. The structures were reported in complex with ligands BIM8 (PDB.1UVR) and 3A3 (PDB.2GCU), respectively. The nonpolar hulls are depicted in yellow, and were computed taking into account only the two PDB complexes.

Only 32 of the 119 assayed kinases are reported in PDB complexes (Fernández and Maddipati, 2006), yet, structural inferences can be made with confidence ((Bonneau *et al.*, 2002; Chivian *et al.*, 2005; Fernández and Berry, 2004) given the kinase homology ((Man-

ning *et al.*, 2002) and high alignment (RMSD <0.9 Å) across reported structures (Fernández and Maddipati, 2006). The prediction accuracy decreases somewhat but in a quantifiable manner ($\sim 12\%$) on loopy regions (section 2.2.3). The matrix $\mathbf{D}_{np} = [d_{np}(i,j)]$ (Figure 3.3)

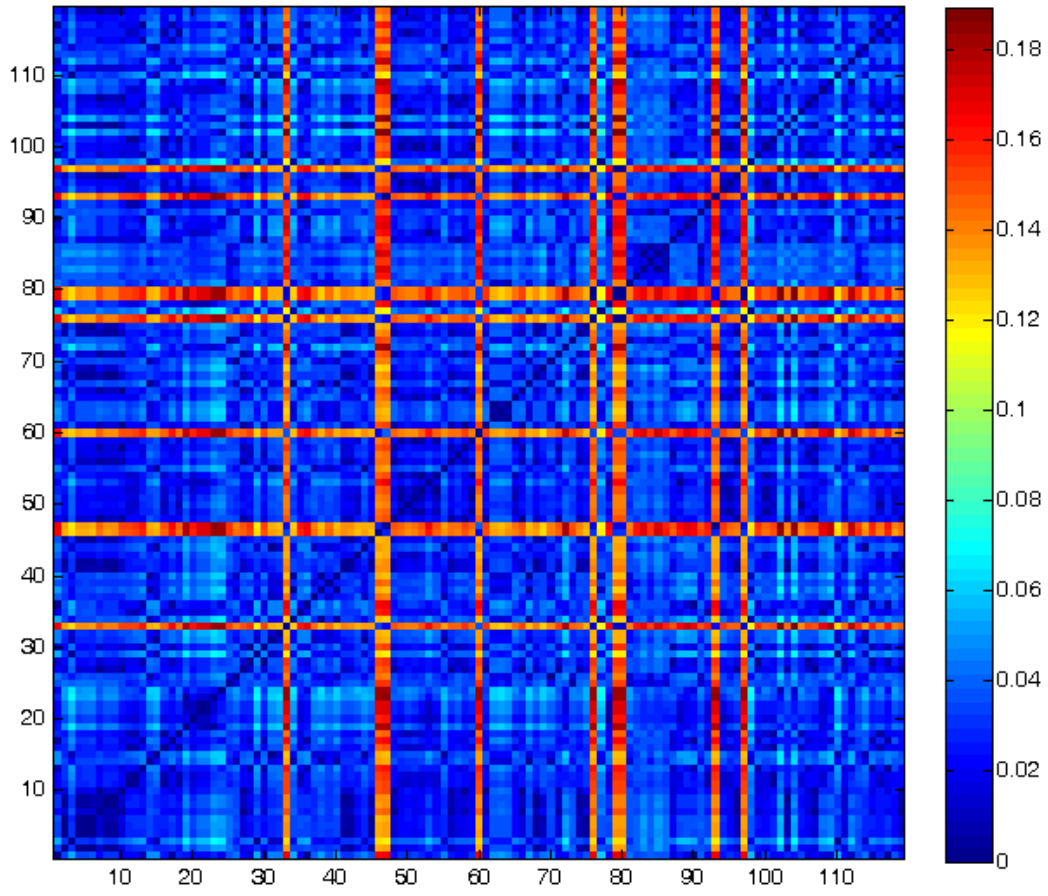


Figure 3.3: Nonpolar distance matrix $\mathbf{D}_{np} = [d_{np}(i,j)]$ over the 119 assayed kinases. The numerals in rows and columns follow Figure 3.1

reveals remarkable nonpolar similarity across kinases: $< d_{np} > /(\max(d_{np})) \approx 11\%$ (<> = average over kinase pairs); $< [d_{np} - < d_{np} >]^2 >^{\frac{1}{2}} / < d_{np} > \approx 16\%$. This similarity im-

plies a high conservation of accessible nonpolar surface, an indication that ligands whose affinity is dominated by hydrophobic interactions should be highly promiscuous.

The plot d_{np} vs d_{phar} for all $(119 \times 118)/2$ kinase pairs (i, j) shows no correlation between the two metrics (Figure 3.4). This suggest that nonpolar pattern should not be

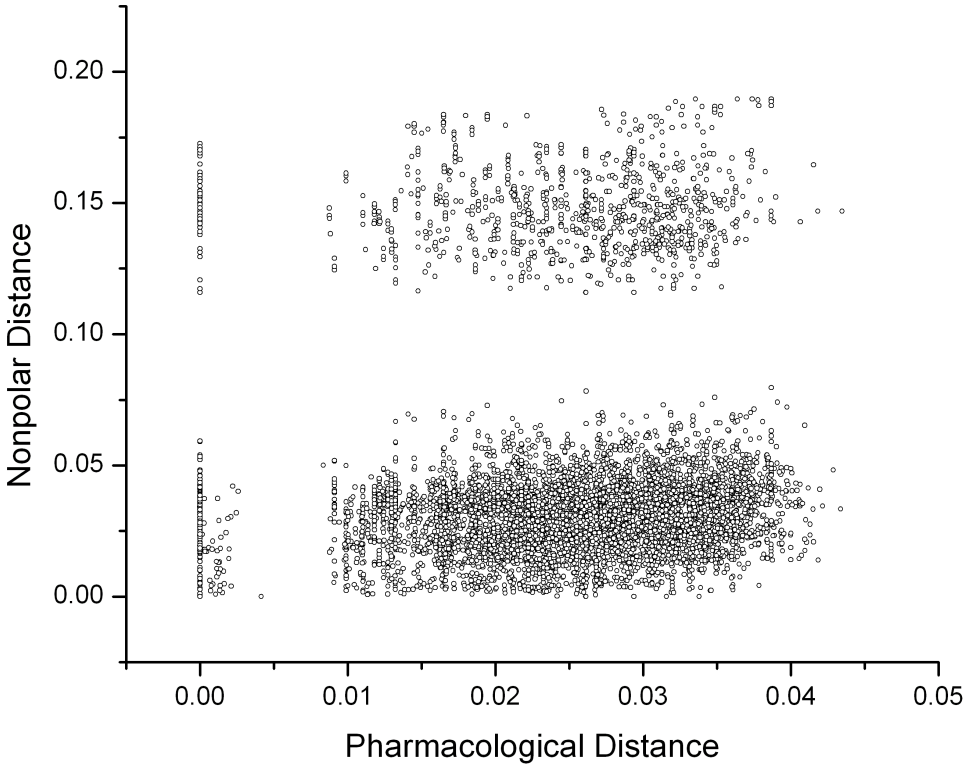


Figure 3.4: Plot of nonpolar distance versus pharmacological distance. Each circle represents a kinase pair. No correlation is observed, while there is some bimodality in each dimension.

the molecular basis for kinase inhibitor. However, when the highly promiscuous affinity-dominant staurosporine is incorporated to the affinity profile (Fabian *et al.*, 2005) and the

affinity-based distance matrix is recalculated ($d_{phar} \rightarrow d_{ps}$ = pseudopharmacological distance), a good correlation ($R^2 = 0.875$) between d_{ps} and d_{np} is obtained (Figure 3.5). This

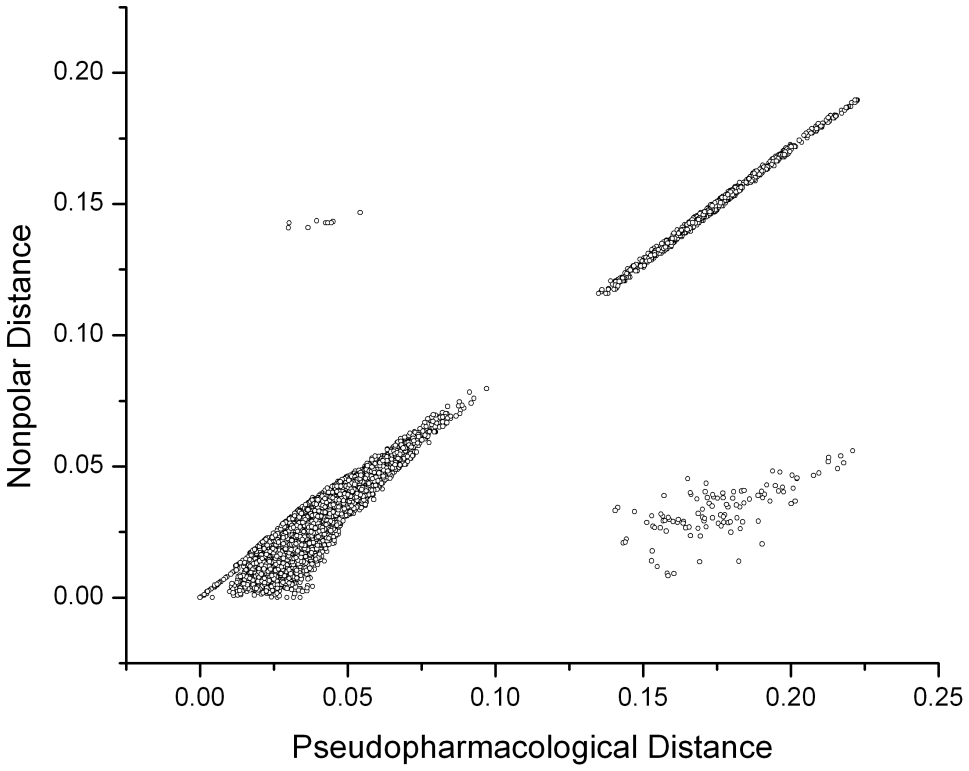


Figure 3.5: correlation between pseudopharmacological distance (including staurosporine in the drug screening background) and nonpolar distance between kinases. The sole outliers are pairs involving the EGFR kinase, the kinase whose affinity vector is not dominated by staurosporine (cf. Fabian *et al.* (2005), Figure 5).

correlation reveals that promiscuity, the dominant affinity trait when staurosporine is incorporated, is fostered by targeting accessible nonpolar moieties, in turn, a highly conserved feature of protein interfaces (Ma *et al.*, 2003). The strong correlation shown in figure 3.5

implies that staurosporine should bind mainly through hydrophobic contacts, as it is indeed the case in its PDB complexes (Fernández and Maddipati, 2006).

Significantly, the most promiscuous drug target, the pregnane X nuclear receptor (PXR, PDB accession code 1ILH), believed to bind to over 50% of human drugs (Hopkins *et al.*, 2006) contains the most extensive nonpolar hull per 1000 Å^2 of ligand surface of all complexes reported in PDB (Figure 3.6), and the highest density of nonpolar accessible surface: 660 Å^2 per 1000 Å^2 of accessible surface within the nonpolar hull.

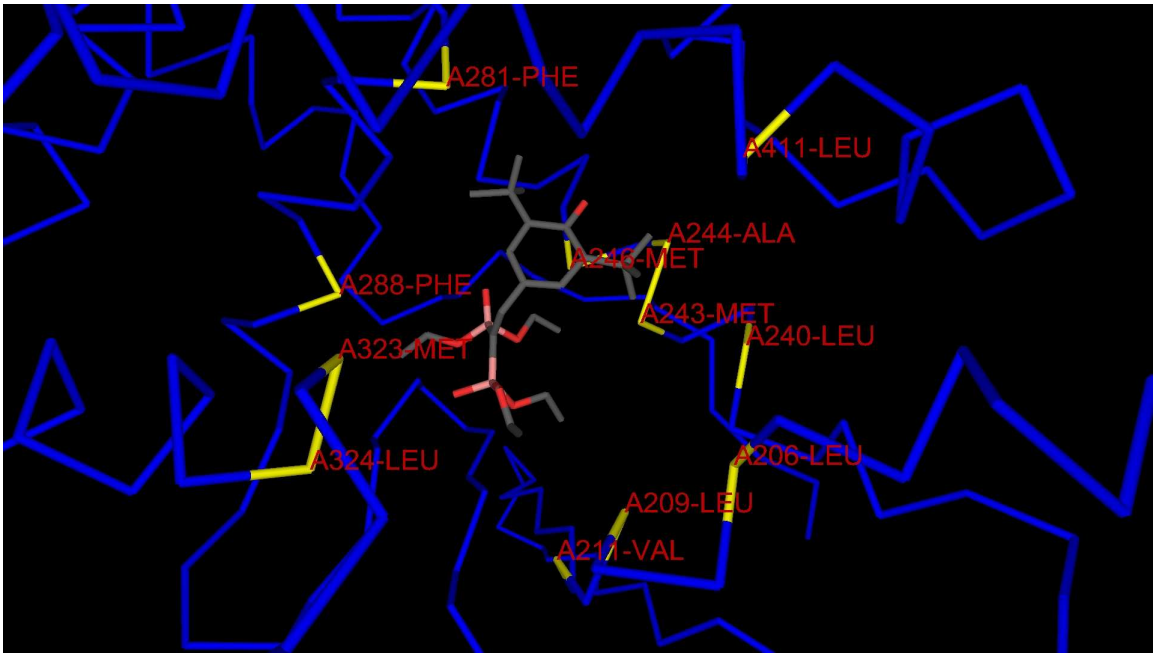


Figure 3.6: The nonpolar hull of the active fold of pregnane X receptor (PXR) in complex with SR12813 (PDB.1ILH). The virtual bonds between α-carbons are depicted in blue, while the residues in the hull are shown in yellow.

On the other hand, sequence alignment including the 32 assayed kinases with PDB-reported complexes (Fernández and Maddipati, 2006) reveals that residues in ATP-binding sites engaged in hydrogen bonding with ligands are highly conserved, with $0 \leq \sigma(n) \leq 0.21$

(n = chain position for hydrogen-bonding residue; σ= information entropy reflecting amino acid variability after sequence alignment); average σ in $H_{np} = 0.87$; $\max(\sigma) = \ln 20 \approx 4.2$ (Higgins *et al.*, 1996; Shenkin *et al.*, 1991). As expected, differences in hydrogen-bonding capabilities do not appreciably correlate with d_{ps} ($R^2 \approx 0.19$), and there is no correlation with d_{phar}.

3.2 Molecular basis for specificity

These observations discussed in the previous section lead to the question: What feature may be targeted to promote specificity? We need to identify a feature with sufficient variability across homologs and capable of influencing the affinity for the ligands by modulating the local propensity for water exclusion. Thus, we focus on "environmental residues", i.e. those residues framing the microenvironment of intramolecular solvent-accessible backbone hydrogen bonds (SAHBs) (Fernández and Berry, 2004). These bonds may become intermolecularly dehydrated upon ligand association. They promote such associations because the enhancement and stabilization of electrostatic interactions overcomes the thermodynamic cost associated with removing the surrounding water molecules that hydrate amide or carbonyl groups (Fernández and Berry, 2004). Environmental residues include the hydrogen-bonded residues themselves. To compare environments of different kinases, we define the environmental hull of a kinase, H_{env}, as the reunion of all environmental residues in the chain and residues aligning with environmental residues from other chains (figure 3.7, see section 2.2.4 for details). Thus, an environmental distance $d_{env}(i, j)$ between kinases i and j is obtained by comparing the aligned hydrogen-bond microenvi-

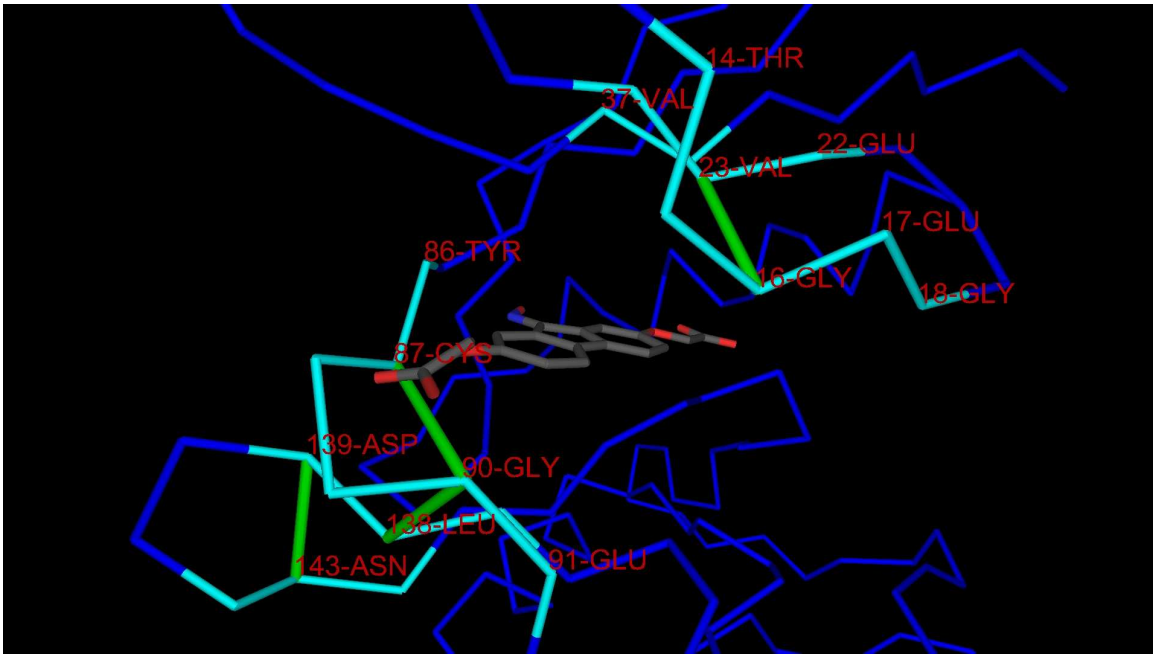


Figure 3.7: Environmental hull (light blue) for CHK1 (obtained from alignment with PDK1). Solvent-accessible hydrogen bonds (SAHBs) are indicated as green segments joining the α-carbons of the paired residues. The virtual bonds are shown as blue segments. The three SAHBs perturbed by the ligand (3A3) are C87-G90; G90-L138; G16-V23.

ronments within $H_{env}(i)$ and $H_{env}(j)$:

$$d_{env}(i,j) = \frac{1}{M(i,j)} \sum_{n=1,\ldots,M(i,j)} \Delta_n(i,j), \tag{3.3}$$

where $M(i,j)$ = number of residue pairs in $H_{env}(i)$ corresponding to SAHBs in kinase i or to hydrogen bonds or nonbonded residue pairs that align with SAHBs in $H_{env}(j)$; n = dummy index denoting residue pair, and $\Delta_n(i,j) = 1$ if residue pair n corresponds to a SAHB in $H_{env}(i)$ that aligns with a non-SAHB in $H_{env}(j)$ or *vice versa*, and $\Delta_n(i,j) = 0$, otherwise. Thus defined, the environmental distance compares local dehydration propensities associ-

ated with SAHB patterns in kinases. The validity of the relation: $\Phi_i(H_{env}(j)) = H_{env}(i)$ enables a rigorous comparison between environments of two different proteins, in contrast with earlier attempts (Fernández and Maddipati, 2006).

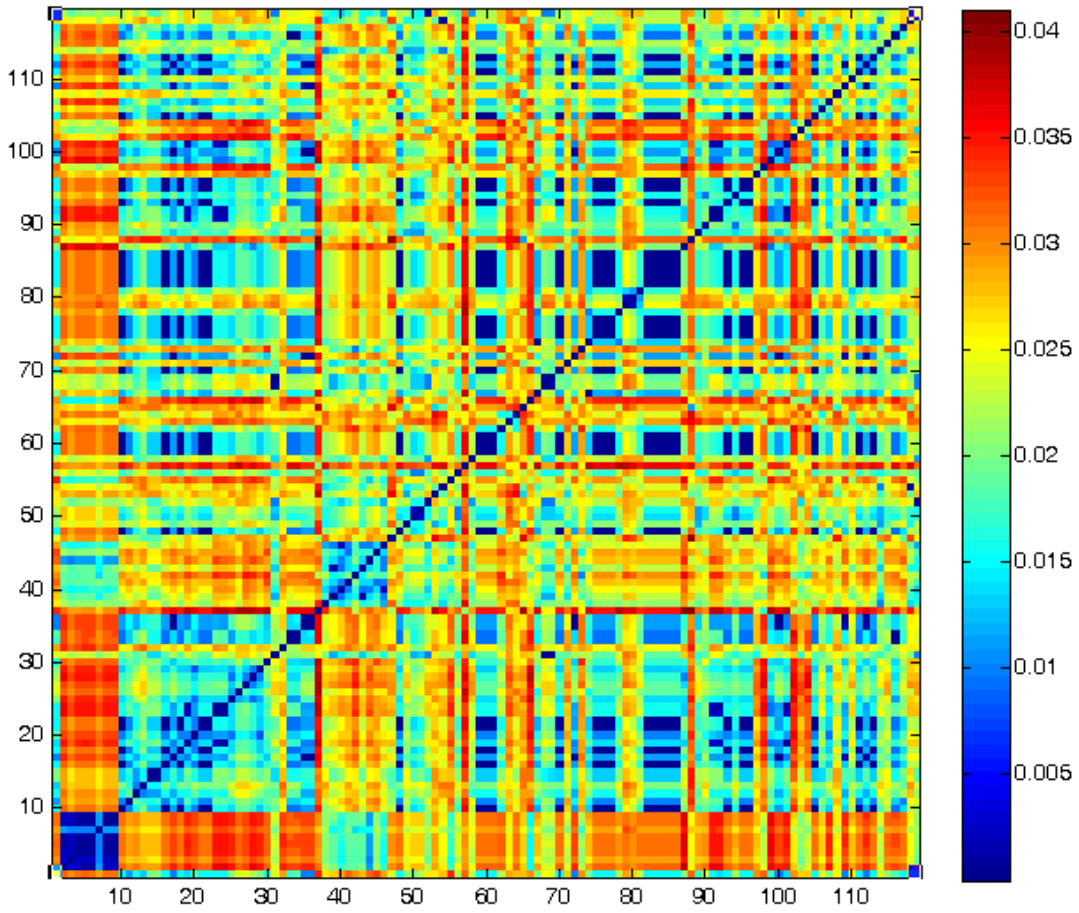


Figure 3.8: Environmental distance matrix $\mathbf{D}_{env} = [d_{env}(i,j)]$ for the 119 kinases assayed (Fabian *et al.*, 2005).

The matrix $\mathbf{D}_{env} = [d_{env}(i,j)]$ for the 119 kinases assayed (Fabian *et al.*, 2005) (figure 3.8) is obtained after inference of the SAHBs for the 87 kinases unreported in PDB from

direct structure prediction (Bonneau *et al.*, 2002; Chivian *et al.*, 2005). The predictions are validated through a correlation with an independent and accurate sequence-based prediction of another structural attribute: the disorder propensity (Braken *et al.*, 2004; Fernández and Berry, 2004). This attribute was chosen because loopy regions, in the twilight between order and disorder, compromise somewhat the accuracy of a structure prediction. The validation is based on the fact that the extent of intramolecular hydration of a hydrogen bond correlates strongly with the disorder propensity: disorder arises from an impossibility to sufficiently hinder hydration of amides and carbonyls (Fernández and Berry, 2004) (see section 2.2.3 for details). The strong correlation between d_{env} and d_{phar} ($R^2 \approx 0.917$) (Figure 3.9) reveals that the impact of drugs on the human kinome is dictated by differences in hydration microenvironments across the ligand-binding regions of the kinases. To the best of our knowledge, the hydration differences across kinases, quantified through the metric d_{env}, were not considered in the development of the drugs screened in Fabian *et al.* (2005). The diversity in hydration microenvironments needed to yield specificity across paralog kinases results from the variability ($<\sigma>\approx 1.38$) of environmental residues, while $<\sigma>\approx 0.21$ when the average is restricted to residues paired by hydrogen bonds that are environmentally affected by the ligands.

The metric d_{env} is defined at the sequence level by identifying the comparison window through environmental alignment (section 2.2.4). This technique required aligning residues with those in homolog kinases with reported structure whose microenvironments are known to be affected by ligand association. Thus, in contrast with the structure-based packing distance that compares SAHBs wrapped by ligands in the structures of two complexes (Fernández and Maddipati, 2006), a background of homologs, S, is needed to com-

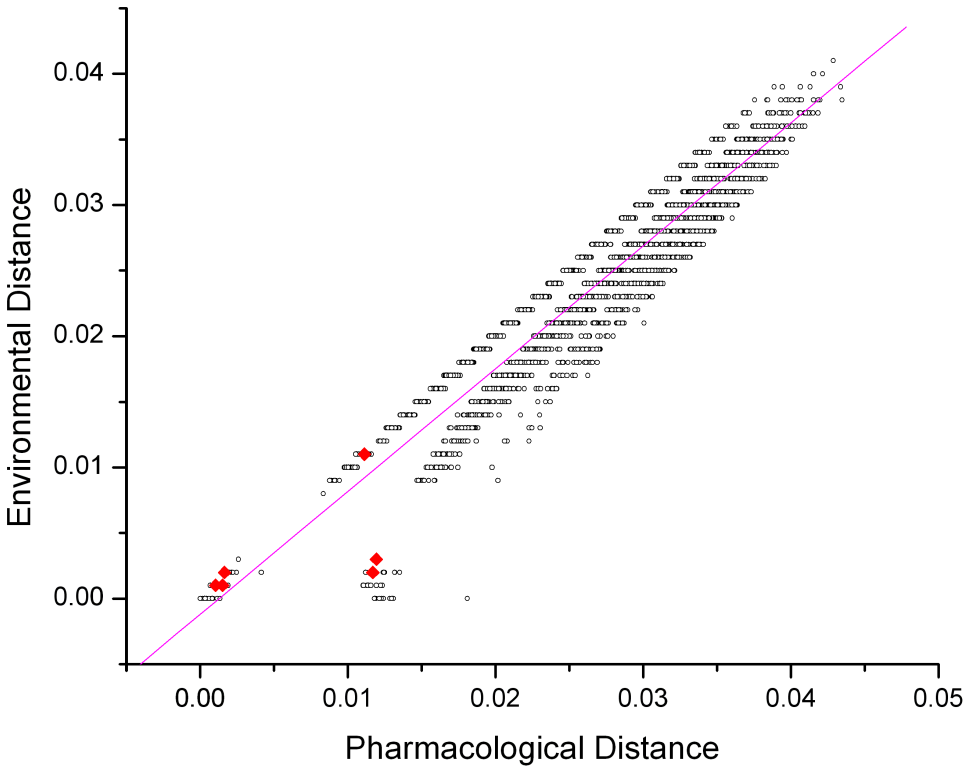


Figure 3.9: Correlation of environmental versus pharmacological distance. The line indicates the optimal linear fit. The red diamonds correspond to the six pairs including ABL1, the primary target for imatinib, and each of its six mutants, listed in Figure 3.1, that confer different degrees of drug resistance.

pare sequence pairs using d_{env}. If S is limited (cardinal <5), d_{env} is well approximated by d_{pack}, but for a more extended and reliable window of comparison (determined by aligning the test sequence to more PDB-homologs), d_{env} becomes a distinctive metric (Figure 3.10). This extended comparison is required to make reliable sequence-based inferences and incorporates for a given protein the alternative binding regions found in homologs.

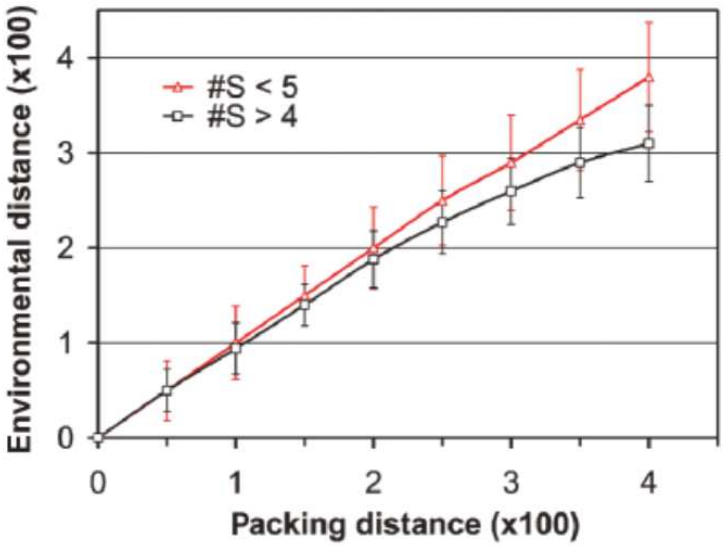


Figure 3.10: Relation between packing and environmental distance as function of the size, #S, of the structural background set used to define the environmental hull. The 103 structurally reported kinases were used for the analysis and their environmental distances were computed as if the structure were unknown. For a reduced background (#$S < 5$), the packing metric is well approximated by d_{env}, although with significant dispersion ($\sim 25\%$, error bars). As more structural background is included (#$S > 4$), packing distance becomes an overestimation.

Local hydration differences determining specificity of drug inhibitors may arise in kinase pairs with high degree of structural alignment (Hogue, 1997) ($RMSD \approx 0.3$ Å), such as SRC and LCK. The latter is a known target for Gleevec (imatinib), which may thus act as immunosuppressant (Dietz *et al.*, 2004), while SRC shows no affinity for the drug (Fabian *et al.*, 2005). Strikingly, two SAHBs in LCK, G254-G257 and R397-A400 that promote their own dehydration through protein-imatinib association (figure 3.11), are absent from SRC: regions in the twilight zone between order and disorder in LCK become fully disordered in SRC, as shown by the absence of electron density (figure 3.11). Thus,

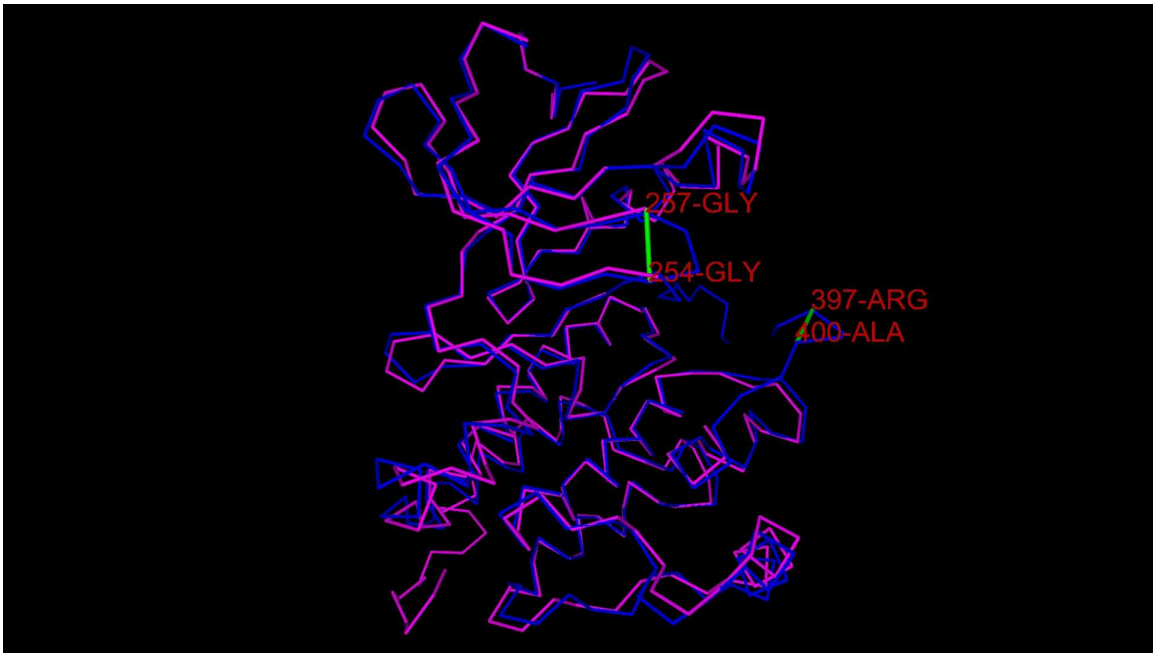


Figure 3.11: Environmental differences between the highly alignable native folds of LCK (blue) and SRC (lilac). The two SAHBs G254-G257 and R397-A400 are present only in LCK, a target for imatinib, while SRC has no affinity for the ligand.

while structurally similar, the environmental distance between SRC and LCK is sufficient to account for the selective affinity of imatinib.

Drug-resistant mutations in another imatinib target, Bcr-ABL (ABL1), produce environmental differences that correlate approximately with the pharmacological distances between mutant and wild type (red diamonds in figure 3.9). Upon close examination, the most effective mutations (Fabian *et al.*, 2005) T315I and E255K are precisely the ones that have the most dramatic effect in increasing the dehydration (by adding nonpolar groups to the microenvironment) of preformed SAHBs Q300-E316, and G251-G254, respectively. These SAHBs are part of the environmental hull of wild-type Bcr-ABL. All reported ABL1 mutations actually perturb the dehydration propensity of SAHBs (figure 3.12). The poor-

est correlation between environmental and pharmacological distance (figure 3.9) arises for mutations E255K, H396P, likely to perturb affinity through other mechanisms that supersede environmental change. Overall, the drug-resistant mutations significantly decrease the dehydration propensity of the target surface and accordingly decrease the inhibitor affinity.

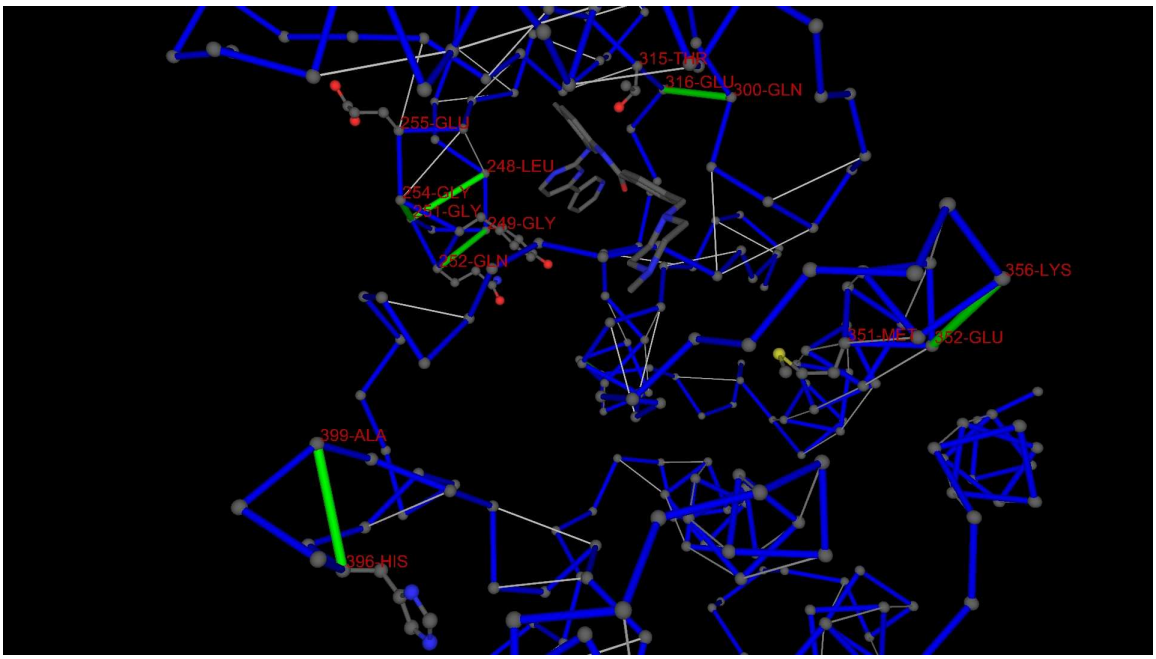


Figure 3.12: Environmental impact of the drug-resistant mutations of ABL, a primary target for imatinib (PDB.1IEP, ligand shown in complex). Only the side chains of the mutating residues are indicated, together with the SAHBs (green) whose microenvironments they affect. Hydrogen bonds not accessible to solvent are shown as thin segments in light grey. The mutations with the SAHBs affected (in brackets) are: T315I (Q300-E316); E255K (G251-G254); Q252H (L248-G251; G249-Q252; G251-G254); Y253F (L248-G251; G249-Q252; G251-G254); M351T (E352-K356); H396P (H396-A399).

3.3 Conclusion

In this chapter, we identified the molecular code for promiscuity and specificity in targeting drugs that impact the human kinome. Thus, significant progress in the informatics of drug design has been achieved by determining the type of similarity among targets that promotes promiscuity and the type of difference that controls specificity. The former corresponds to highly conserved exposed nonpolar moieties in alignable regions of protein-ligand interfaces, while the latter corresponds to nonconserved hot spots of high dehydration propensity around amide-carbonyl pairs.

Kinase structures may be tightly aligned except in the regions of highest conformational plasticity, precisely the regions of catalytic and regulatory relevance (Huse and Kuriyan, 2002). These loopy regions are obviously rich in packing defects, as their backbone hydrogen bonds are not fully buried and thus contain hot spots of dehydration propensity. Thus, a way to achieve a classification of kinases is to assess differences in the nonconserved microenvironments of their loopy regions. As shown in this thesis, such microenvironmental differences reflect the differences in binding affinity against a representative set of drugs.

The analysis was carried out at a genomic scale, since the environmental assessment of local dehydration propensities can be reliably determined from sequence (c.f. Fernández and Berry, 2004). The sequence-based predictor became operational because of the high level of structural similarity reported for the kinase superfamily, combined with relatively low sequence homology levels that enable target discrimination. These properties paved the way to the mapping of a reliable sequence-based attribute, the disorder propensity, onto a molecular feature, the dehydration propensity, shown to be effectively sculpted in drug

ligands to modulate specificity.

In future work, our methodology will be adapted to model specificity control in drug-targeted nuclear hormone receptors. The moderate structure conservation of the ligand-binding domain (Escriva *et al.*, 2000) and the significant PDB representation needed to construct a reliable structural background instills confidence in the success of our environmental predictor.

The problem in this chapter is essentially handled by calculating pharmacological distances (distances between vectors) from affinity profiles (vectors). However, in practical work it is the affinity profiles that drug developers are finally concerned with. This leads us to the inverse problem of determining affinity profiles of individual kinases from their predicted pharmacological distances. This problem will be investigated in next chapter.

Chapter 4

In silico drug profiling of the human kinome

In Chapter 3 we identified the molecular basis for drug promiscuity and specificity. In this chapter we introduce an application of this discovery of the structural marker governing specificity. This application is to solve one of the major problems in drug development, i.e., the screening of drug profile against a large protein library. Let's start with a brief review regarding the background of the problem: affinity-profile screening in drug development.

A primary problem in drug development, to identify compounds with controlled specificity against clinically relevant targets, is usually handled in early stages of development by high-throughput screening, both experimentally and computationally (*in silico*)(Drews, 2000; Bleicher *et al.*, 2003). Due to the increasing cost resulting from high clinical failure rates in a downstream stage of development, drug designers are prone to terminate the efforts on those compounds likely to fail in late stages as early as possible(Bleicher

et al., 2003). This "fail early" strategy places a significant responsibility on the early-stage compound profiling(Liszewski, 2006). The two basic types of screening, experimental and computational, are complementary approaches: the former is relatively accurate but cost-limited by the size of the compound library to be profiled (Fabian *et al.*, 2005), while the latter is more time- and cost-efficient but less reliable(Bleicher *et al.*, 2003; Kitchen *et al.*, 2004; Shoichet, 2004). *In silico* screening methods(Oprea and Matter, 2004) can be categorized as either ligand-based(Lengauer *et al.*, 2004) or target-based(Lyne, 2002). The latter is becoming mainstream for cases where target structural information is available(Kitchen *et al.*, 2004; Oprea and Matter, 2004; Kuhn *et al.*, 2002). Most of the target-based virtual screening is performed by docking and scoring(Lyne, 2002). However, the inaccuracies of scoring functions pose a major problem in target-based virtual screening(Kitchen *et al.*, 2004; Shoichet, 2004). Such docking-based algorithms are inadequate to examine kinase targeting(Huse and Kuriyan, 2002; Mizutani and Itai, 2004; Mizutani *et al.*, 2006; Chen *et al.*, 2007; Bain *et al.*, 2003; Druker, 2004; Hopkins *et al.*, 2006; Knight and Shokat, 2005; Vieth *et al.*, 2004) because most of them do not take into account the induced fits upon binding, which are crucial for kinase-ligand associations that typically involve loopy regions(Huse and Kuriyan, 2002). There exist some docking-based algorithms that take into account induced fits, but they still cannot handle extensive induced fit adaptation including large movements of the backbone(Mizutani and Itai, 2004; Mizutani *et al.*, 2006), which is just the case for the loopy regions in kinases. In fact, kinase ATP-pockets are partly framed by floppy regions, including the activation loop, catalytic loop and P-loop(Huse and Kuriyan, 2002). These parts of the structure undergo an order-upon-binding transition upon association with the ligands (natural or otherwise) which cannot be captured with a dock-

ing algorithm, no matter how well it is able to handle flexibility. The induced fit problem as it stands today is almost as hard as the protein folding problem and no algorithm deals with it effectively from first principles.

This problem notwithstanding, the molecular markers for specificity and promiscuity discussed in Chapter 3 may herald the advent of novel predictive tools for drug affinity profiling by providing a new vantage point for kinase comparisons, as described in this chapter. These comparisons will enable us to predict cross reactivities. Unexpected cross reactivities became recently apparent with the advent of high-throughput experimental screening techniques(Fabian *et al.*, 2005) based on bacteriophage kinase display. Thus, the affinity profiles of 20 inhibitors against a battery of 119 kinases have been reported(Fabian *et al.*, 2005). However, the operational value of these assays to identify leads from within a compound library (1000 compounds) is limited by cost, justifying our development of *in silico* profiling tools.

In this chapter we introduce a predictive profiler based on the assumption that a structure-based feature-similarity comparison of molecular targets can be used as a surrogate for the differences in their pharmacological behavior. Our approach exploits the the discovery of the structural marker governing specificity discussed in Chapter 3, a result holding even for kinases lacking PDB representation and, based on this feature, it introduces a comparison of kinases including purported targets and experimentally confirmed targets. The core of our affinity-profile predictor involves determining a linear propagator of profiling data. This propagator consists of the structure-based estimation of pharmacological distances across kinases. Once the propagator is computed, the inference of affinity profiles for test drugs becomes a problem in distance geometry.

4.1 Procedures of Drug Profiling

4.1.1 Operational premises of the predictor.

In Chapter 3, we introduced the environmental distance matrix ($\mathbf{D}_{env}$) for all kinase pairs obtained from the respective differences in the SAHB (i.e., dehydron) patterns within aligned structures. On the other hand, the differences in the affinity profiles of kinases against a background of drugs are quantified and arrayed in the so-called pharmacological distance matrix $\mathbf{D}_{phar}$. Furthermore in Chapter 3, we revealed a tight linear correlation between the environmental and the pharmacological distances (c.f. Section 3.2), thus delineating the molecular basis for cross reactivity. The profiling method presented in this chapter is an application of this correlation to estimate $\mathbf{D}_{phar}$ from $\mathbf{D}_{env}$ and further, to construct a full affinity profile for a test compound from its sub-profile obtained from experimental affinity assays against a small kinase subset. This problem becomes a linear algebra problem, where the full affinity vector for a given drug may be determined from suitably defined distances between kinases and the experimentally obtained affinity vector.

The implementation of the profiling method is thus carried out according to the following steps (sketched in Figure 4.1 and Figure 4.2):

1. Estimate the pharmacological distance matrix ($\mathbf{D}_{phar}$) defined over all kinase pairs by appropriately rescaling the environmental distance matrix ($\mathbf{D}_{env}$). This estimation will increase in accuracy as a more comprehensive background of drugs, i. e. one covering all kinase targets, is used to determine pharmacological behavior.

2. For every test compound to be fully profiled in silico, choose a subset of kinases to

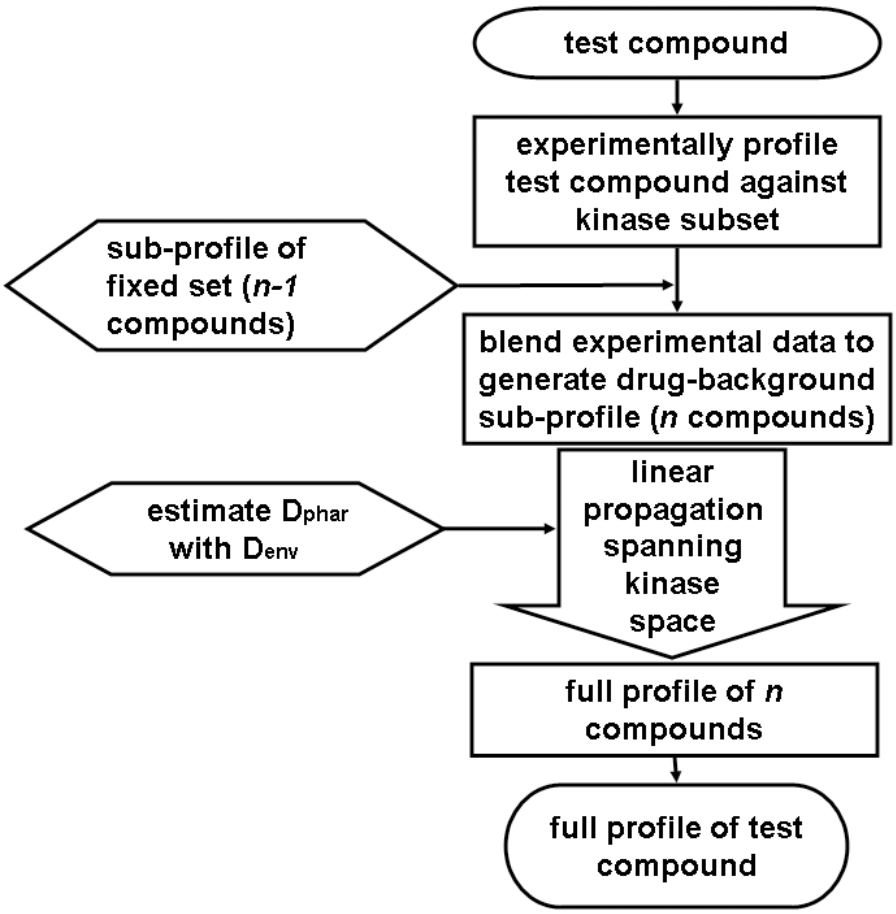


Figure 4.1: Flow chart (left) of the *in silico* profiling method.

obtain a small-scale experimental affinity profile, heretofore noted the "sub-profile". The required size of the kinase subset sampled to produce the sub-profile depends on the fixed number of inhibitors that define the pharmacological matrix $\mathbf{D}_{phar}$.

3. Determine the full affinity profile for a test compound from its sub-profile and our estimated $\mathbf{D}_{phar}$.

4. Repeat the steps above for each test compound in the library.

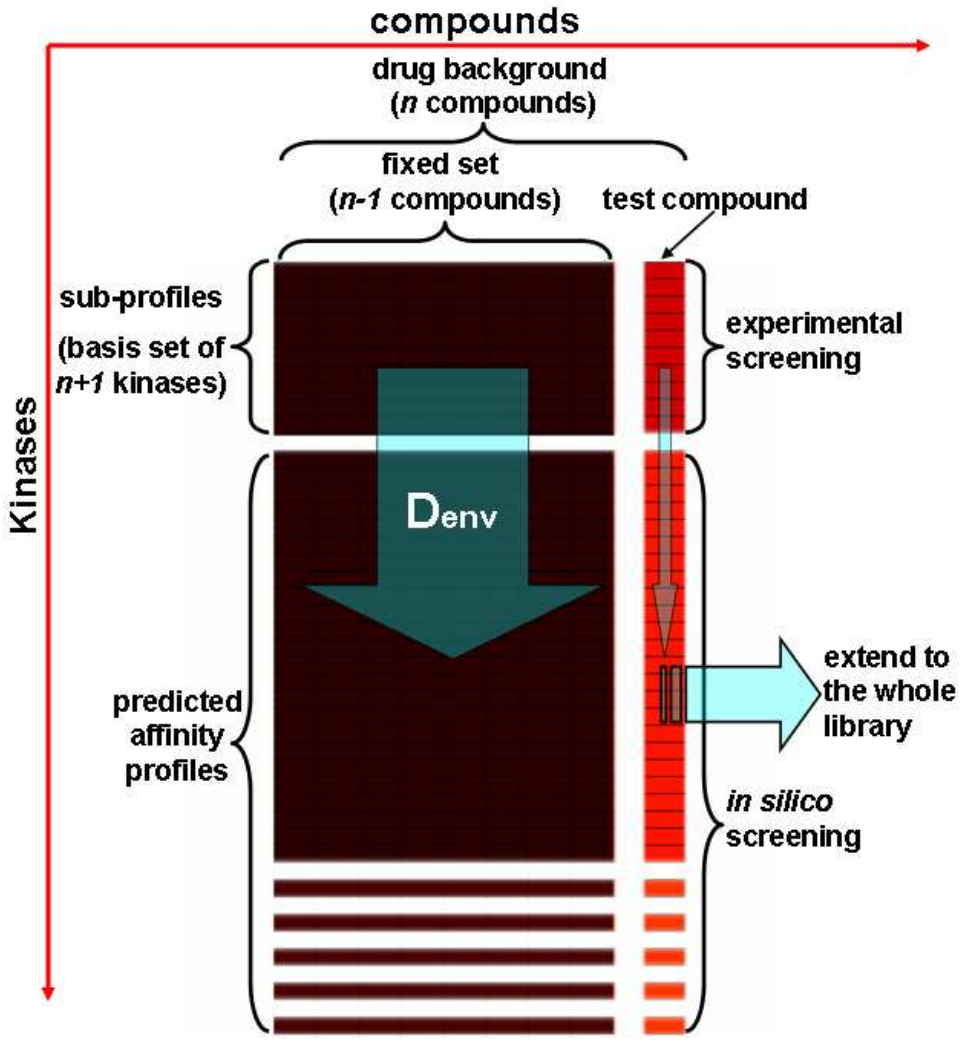


Figure 4.2: Process diagram (right) of the *in silico* profiling method. Each column corresponds to one compound and each row to one kinase. The brown columns correspond to the compounds with profiles already known and the red column corresponds to the test compound. The upper rows represent the sub-profiles that are obtained from experiments and the lower rows represent the profiles predicted by the profiler.

The sub-profile used as input information in Step 2 and Figure 4.1 and 4.2 is obtained

from small-scale experiments, most advantageous in the case when a battery of a large number of kinases needs to be screened.

4.1.2 Estimating pharmacological distances from environmental distances

To infer $\mathbf{D}_{phar}$, it becomes necessary to identify the structural feature that governs drug cross reactivity across paralog kinases. It is revealed in Chapter 3 that SAHBs constitute such structural markers. Thus, we estimate $\mathbf{D}_{phar}$ from $\mathbf{D}_{env}$ obtained by comparing the SAHB patterns of purported targets. The matrix $\mathbf{D}_{env}$ quantifies differences in the SAHB patterns within the ATP pockets (i.e. the drug binding site) across all kinase pairs. Thus, the environmental distance is based on structural alignment followed by comparison of poorly conserved features. The determination of SAHBs is following the procedures in Chapter 3.

In Section 3.2 we already showed the correlation between $\mathbf{D}_{phar}$ and $\mathbf{D}_{env}$ when the former is obtained from the T7-bacteriophage kinase display screening against a background of 17 drugs(Fabian *et al.*, 2005). For each kinase pair (i, j), we can then infer the pharmacological distance using the linearly fitted parameters. Figure 3.9 is plotted with normalized pharmacological distances. For the purpose of inference, here we replot the correlation with original pharmacological distances (Figure 4.3) and use the fitted parameters based on such data:

$$\mathbf{D}_{phar}(i,j) = 1985.5\mathbf{D}_{env}(i,j) + 7.0307 \tag{4.1}$$

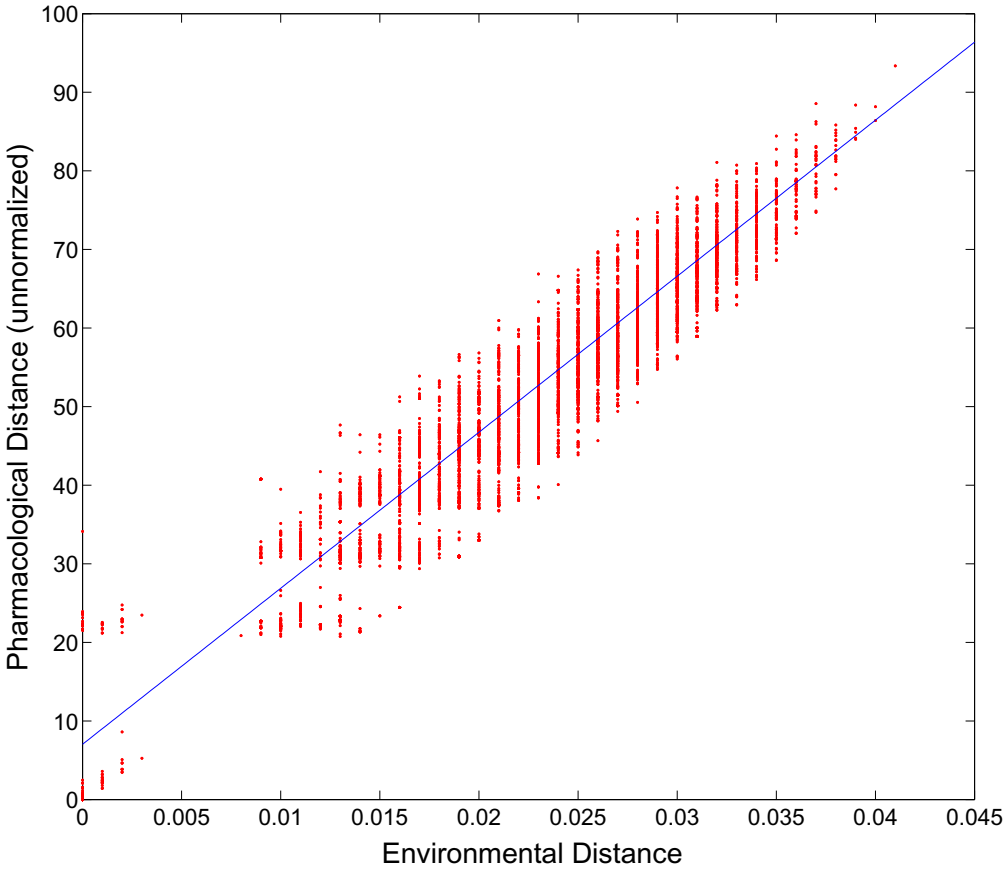


Figure 4.3: Correlation between environmental and pharmacological distances. Each diamond represents a pair of kinases with horizontal coordinate being the environmental distance and vertical coordinate being the pharmacological distance between them. Unlike Figure 3.9, the environmental distances in this figure are not normalized. The straight line indicates the linear fit by least-squares method: $\mathbf{D}_{phar}(i,j) = 1985.5\mathbf{D}_{env}(i,j) + 7.0307$. Note that the correlation is very tight ($R^2 \sim 0.92$).

In the process of inferring $\mathbf{D}_{phar}$ from $\mathbf{D}_{env}$, some errors are introduced in the estimated pharmacological distances. There are at least two possible sources of errors leading to dispersion in the correlation:

a) The background of drugs used to define the pharmacological distance is limited, with uneven target coverage, and thus only approximately indicative of pharmacological behavior.

b) The SAHBs are not the only selectivity determinant for nonpromiscuous drugs.

In regards to error source (a), we are limited in our analysis by the availability of drugs chosen to define pharmacological profiles in high-throughput experiments (Fabian *et al.*, 2005). We may need to revise $\mathbf{D}_{phar}$ as new screening data becomes available. In regards to (b), we can only claim that SAHBs are one determinant but not necessarily the only factor governing ligand specificity (Chen *et al.*, 2007). Even though these errors will be inherited in the following steps, we made the method less sensitive to systematic errors through adequate parameterization.

4.1.3 Expanding pharmacological information from limited affinity profiles

We now show how to determine the affinity profiles of kinases from structure-based estimations of pharmacological distances between kinase pairs. Just like vector coordinates cannot be uniquely determined from vector distances, affinity profiles cannot be uniquely determined solely from the pharmacological distances: additional constraints are required. To solve the profile prediction problem, we cast it in terms of linear algebra, through a (vector) $\leftrightarrow$ (kinase profile) correspondence. The procedure boils down to determining vector coordinates given distances between vectors. To guarantee the uniqueness of the solution, a subset of vectors should be given in addition to vector distances. Thus, for given

pharmacological distances between pairs of kinases, the number of independent degrees of freedom of the solution vectors (kinase profiles) is n, which corresponds to the number of inhibitors, i.e., the dimension of the affinity-profile space. Therefore, we need to know at least n independent vectors to determine all affinity profiles. These conditions narrow down solutions to two possibilities, due to symmetry. Reflection relative to the hyperplane determined by the fixed n independent vectors produces two conjugate solutions to the problem (see the example in Figure 4.4). Thus, to unambiguously determine the solution, we need the coordinates of an additional vector. Accordingly, the minimal number of fixed vectors should be $n+1$. A constraint on these $n+1$ vectors is that n of them should be linearly independent, i.e., they should not be within one (n-1)-dimensional hyperplane.

To summarize, we may re-cast the profile prediction problem in linear-algebra terms through the following correspondences:

Space dimension $\leftrightarrow$ Number of sampled inhibitors

Distance $\leftrightarrow$ Pharmacological distance

Vector $\leftrightarrow$ Affinity profile of kinase (against a background of inhibitors)

Vector $\mathbf{x}_i \leftrightarrow$ Experimentally determined profile of the i^{th} kinase in the subset

Vector $\mathbf{y} \leftrightarrow$ Affinity profile of the test kinase

Thus, the *in silico* profiling problem may be cast as a linear-algebra problem as follows: In an n-dimension space, we have $n+1$ given vectors (the subset) with n of them being linearly independent:

$$\mathbf{x}_i, \quad i = 0, 1, 2, ..., n \tag{4.2}$$

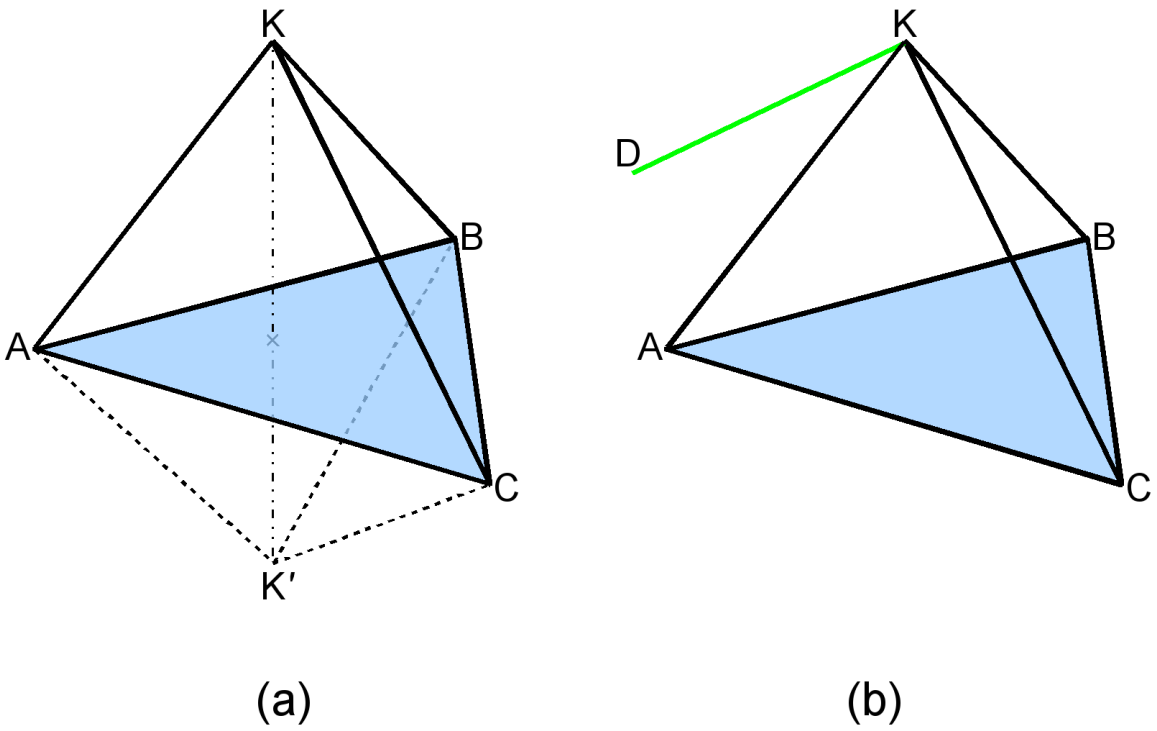


Figure 4.4: A 3-dimensional example illustrating the neccessary conditions to uniquely determine a system of points. **(a)** In 3-dimensional space, a group of points with all the distances between them are given. If only the coordinates of three points (A, B and C) are provided, then there are two possible cases satisfying the conditions, which are are symmetric to each other with respect to the plane determined by the three given pionts, A, B and C. **(b)** If the coordinates of one more point D that is not in the A-B-C plane are provided, then the conditions are enough to unambiguously determine the solution.

Note that $\mathbf{x}_i$ is the profile vector of kinase i against the n inhibitors. That is, x_i^k equals to the negative logarithm of the dissociation constant between kinase i and inhibitor k. For a generic vector $\mathbf{y}$ (affinity profile of a test kinase) that is not in the subset, we have estimated

the distances d_i from $\mathbf{y}$ to all $\mathbf{x}_i$'s:

$$d_i = |\mathbf{y} - \mathbf{x}_i|, \;\; i = 0, ..., n \tag{4.3}$$

Our purpose is to determine the coordinates of vector y based on the conditions given above. Note that y is the profile vector of a kinase that is not in the subset to be screened by experiment, i.e., y is one of the profile vectors to be determined *in silico*. We now show that this can be achieved by linear algebra calculation.

Since n of the $n+1$ vectors are linearly independent, we can assume without loss of generality that $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are the independent vectors. We first define $\mathbf{x}'_i = \mathbf{x}_i - \mathbf{x}_0$, for $i = 1, 2, \cdots, n$. Note that all $\mathbf{x}'_i$ are linearly independent. Similarly, we define $\mathbf{y}' = \mathbf{y} - \mathbf{x}_0$. Note that now we only need to determine $\mathbf{y}'$ to obtain $\mathbf{y}$. We can determine the scalar product of $\mathbf{y}'$ with each $\mathbf{x}'_i$ using the information on the vector distances and the well known relation $|\mathbf{y}' - \mathbf{x}'_i|^2 = |\mathbf{y}'|^2 + |\mathbf{x}'_i|^2 - 2\mathbf{y}' \cdot \mathbf{x}'_i$, in turn obtained from the definition of scalar product:

$$\begin{aligned} \alpha_i &= \mathbf{y}' \cdot \mathbf{x}'_i \\ &= (|\mathbf{y}'|^2 + |\mathbf{x}'_i|^2 - |\mathbf{y}' - \mathbf{x}'_i|^2)/2 \\ &= (|\mathbf{y}'|^2 + |\mathbf{x}'_i|^2 d_i^2)/2 \end{aligned}$$

where the last expression indicates how α_i can be evaluated from the estimated distances.

We then have a system of n-variable linear equations by rewriting the relationships above

$$\begin{aligned} \mathbf{x}'_1 \cdot \mathbf{y}' &= \alpha_1 \\ \mathbf{x}'_2 \cdot \mathbf{y}' &= \alpha_2 \\ \cdots &\quad \cdots \\ \mathbf{x}'_n \cdot \mathbf{y}' &= \alpha_n \end{aligned} \tag{4.4}$$

Thus the pharmacological problem of determining the profiles of test kinases boils down to solving the system of linear equations (4.4), to obtain $\mathbf{y}'$ and thereby $\mathbf{y} = \mathbf{y}' + \mathbf{x}_0$.

Note that each test-kinase profile $\mathbf{y}$ is represented as a row in the "predicted affinity profiles" in the scheme shown in Figure 4.2.

This calculation can be repeated for each test kinase not included in the subset of experimentally screened kinases. By calculating the "predicted affinity profiles" row-by-row, we then extend the sub-profile to an entire affinity profile over all kinases for which structural information is available (and hence pharmacological distances may be estimated).

4.1.4 Prediction of affinity profiles

In order to find the affinity profile of a new inhibitor, we first obtain its affinity sub-profile against the kinases in the subset. We treat the inhibitor as one of the linear dimensions of the pharmacological distance space. As shown above, we calculate the affinities of all the inhibitors, including the new one, towards the kinases that are not within the subset. In this way, the entire profile of the new inhibitor is obtained.

In an ideal case as a mathematical model, if the estimated $\mathbf{D}_{phar}$ were highly close

to the real values, we could obtain the quantitatively exact profile (i.e., exact values of K_d's or ΔG's). Unfortunately, the correlation of $R^2 \sim 0.92$ is still not tight enough for a quantitative profile prediction. Alternatively, we predict the profile in a qualitative level, i.e., "hit" or "no hit". At this stage, we use thresholds of the ligand-target dissociation constant ($K_{d,threshold}$) to determine whether an inhibitor hits a specific kinase or not. That is, if for a kinase-inhibitor pair the predicted K_d is smaller than $K_{d,threshold}$, then we make a qualitative predicttion that the inhibitor binds to the kinase, i.e., a "hit". Three thresholds were used in our predictions: $K_{d,threshold} = 1\mu M, 10\mu M, and 100\mu M$.

Repeating these steps on all test compounds, a large library profiling can be generated from a small-scale (i.e. sub-profile size) experiment. The process is schematically represented in Figure 4.1 and Figure 4.2.

4.1.5 Finding the optimal basis set

The choice of kinases in the basis set (the subset) is crucial for the predictor's performance. First of all, a required condition for the subset is that the affinity vectors corresponding to the kinases span all dimensions of the affinity space. Furthermore, some of the kinases discriminate the inhibitory compounds less effectively than the others, i.e., compounds' affinities to them are more uniform than that to others. If such kinases are included in the subset, the prediction would be more sensitive to errors in the estimated pharmacological distances. This would yield larger errors in the predicted affinity vectors. We then designed a simple algorithm to examine and optimize a series of subset choices. First, the predicted results are benchmarked against the experimental results (Fabian *et al.*, 2005) and the accuracies (percentages of the correct predictions) are calculated. We compared our

predicted results on 17 inhibitors out of the 20 in the screening experiment(Fabian *et al.*, 2005) excluding three promiscuous inhibitors (Staurosporine, EKB-569 and SU11248), against 119 kinases. For the 17×119 entries, we counted the number of entries correctly predicted, in the "hit or no hit" level, and calculated the percentage of correct predictions. Using this percentage as scoring function of the subset-choice, we performed optimizations for the basis set and found several sets with correct prediction percentages around 93%. Our algorithm to optimize the basis set is shown in the pseudo code below:

FOR each kinase in the basis set,

- replace this kinase with each kinase not in the set;
- predict the profile;
- benchmark the predictive results and calculate the accuracy;
- IF the current percentage is higher than the existing best one;
  THEN use this kinase in the place of the original kinase in the set;

END LOOP

Note that this algorithm does not guarantee an optimal basis but only provide a nearly optimal one. The real optimal basis set is extremely hard to find and may only improve the accuracy of the prediction by one or two percentage points from the nearly optimal ones.

4.2 Validation of the Profiler

4.2.1 Experimental validation of affinity predictions

The affinity prediction is validated by benchmarking the result for all the nonpromiscuous inhibitory compounds in the phage-display kinase assay against the experimental data(Fabian *et al.*, 2005). The compounds used in the benchmark are: SB202190, SB203580, VX-745, BIRB-796, SP600125, Gleevec, Iressa, Tarceva, ZD-6474, CI-1033, GW-2016, Vatalanib, MLN-518, LY-333531, BAY-43-9006, Roscovitine and Flavopiridol. For randomly chosen kinase subsets prior to any optimization, mostly the accuracies of the predictions are around $80 \sim 90\%$. [1] For instance, consider the randomly chosen subset of kinases:

AAK1, ABL1, CDK2, EGFR, ERBB2, FLT3, GAK, JNK1, KIT, LCK, p38-alpha, PDGFRB, PHKG1, SLK, SRC, STK10, VEGFR2, YES.

Using this subset, the profiler predicts affinities for all 119 kinases independently assayed (Fabian *et al.*, 2005) with 252 false positives and false negatives out of 17×119 predictions, for the affinity threshold $10\mu M$. The corresponding accuracy is 88%. Notice that the 17×119 predictions cover the sub-profiles for the subset and the accuracy is calculated based on all 17×119 predictions. This does not weaken the validation since when predicting for a kinase within the subset, the input distances are obtained independently, i.e. estimated from environmental distances. In other words, the predictor does not discriminate kinases within or outside of the subset. In a real work, it is useless to predict the

[1] Low accuracies (around 60%) are present but rare. These low accuracies arise from the choices of the kinase subsets extremely vulnerable to the errors introduced in previous step. This is also why we need to perform optimizations - to avoid such choices of subsets.

affinities of an inhibitor for kinases within the subset, but here we just include these results for the sake of validation.

The next step is to optimize the subset of kinases to improve prediction accuracy. For the purpose of validation, we perform the optimization with a "leave-one-out" algorithm. That is, to predict the profile of one inhibitor, we apply the optimization algorithm discussed above to the other 16 inhibitors and thereby get an optimized subset of 17 (=16+1) kinases. These 17 kinases plus one more randomly chosen kinase constitute the subset used for the prediction of the inhibitor. In this way, the inhibitor profile to be predicted is left out of the subset optimization process, and thus the optimized choice of the subset is completely independent of the inhibitor left out, though it might not be the best choice since one kinase in the subset is chosen randomly. It is possible that the randomly chosen kinase plus the 17 optimized ones do not constitute a complete subset (dimension < 17), rendering equation 4.4 not solvable. In this case, we simply replace the randomly chosen kinase with another random one until the corresponding equation 4.4 becomes solvable. As an example, the optimized subset used to predict the profile of SB202190 is as follows (the first 16 kinases result from optimization while the last one is randomly chosen):

INSR, PTK6, CSNK1G1, TEK, Aurora2, EPHA3, PHKG2, BIKE, EPHB4, FYN, TNIK, DAPK2, FGFR1, SLK, PRKAA1, ERBB2, YES, p38-beta.

We predict the complete affinity profiles of the 17 inhibitors one by one, using the profiling information of the 17 inhibitors against the 18-kinase subset. Before predicting for each inhibior, a subset optimization based on the other 16 inhibitors is carried out. That means the subset used in the profile prediction for each one of the 17 inhibitors is not necessarily the same in each case.

The accuracy of our predictor, as revealed by Figure 4.5, is quantitatively summarized in Table 4.1 . In accord with Figure 4.5 , we use three thresholds for kinase-inhibitor hit/no-hit results: $K_{d,threshold} = 1\mu M, 10\mu M, and 100\mu M$. The accuracy of the predictions are 91%, 93% and 93%, respectively. When using $K_{d,threshold} = 1\mu M$, the errors are mainly due to false positives, while for $K_{d,threshold} = 10\mu M$ and $100\mu M$ they arise mostly from false negatives. Thus we can choose different $K_{d,threshold}$'s for different purposes. If we are concerned with drug specificity and promiscuity in a clinical context, it is reasonable to use $K_{d,threshold} = 1\mu M$ since it imposes a more stringent criterion for affinity, suitable for clinical purposes. If binding is the sole concern, a better choice would be $K_{d,threshold} = 10\mu M$ or $100\mu M$, since these filters entail higher sensitivity.

Table 4.1: Prediction accuracy with different filters for hit/no-hit

$K_{d,threshold}$	$1\mu M$	$10\mu M$	$100\mu M$
False positive[†]	116	9	36
False negative[‡]	69	133	114
Accuracy	91%	93%	93%

[†] False positive refers to "no hit" predicted as "hit". [‡] False negative refers to "hit" predicted as "no hit".

The performance and confidence of our predictor are affected by two factors: error sources and the predictor's vulnerability to errors. The major error source in this algorithm arises from the computational estimation of kinase pharmacological distances from environmental distances, which is determined by the tightness of the correlation. Thus, the predictor may perform not so well in the cases where the profile of the compound to predict is not subject to the correlation. The reasons why the correlation is not perfect and the factors affecting the correlation have already been discussed in Section 4.1.2 and may include alternative structural features nonconserved across paralogs. Here we dis-

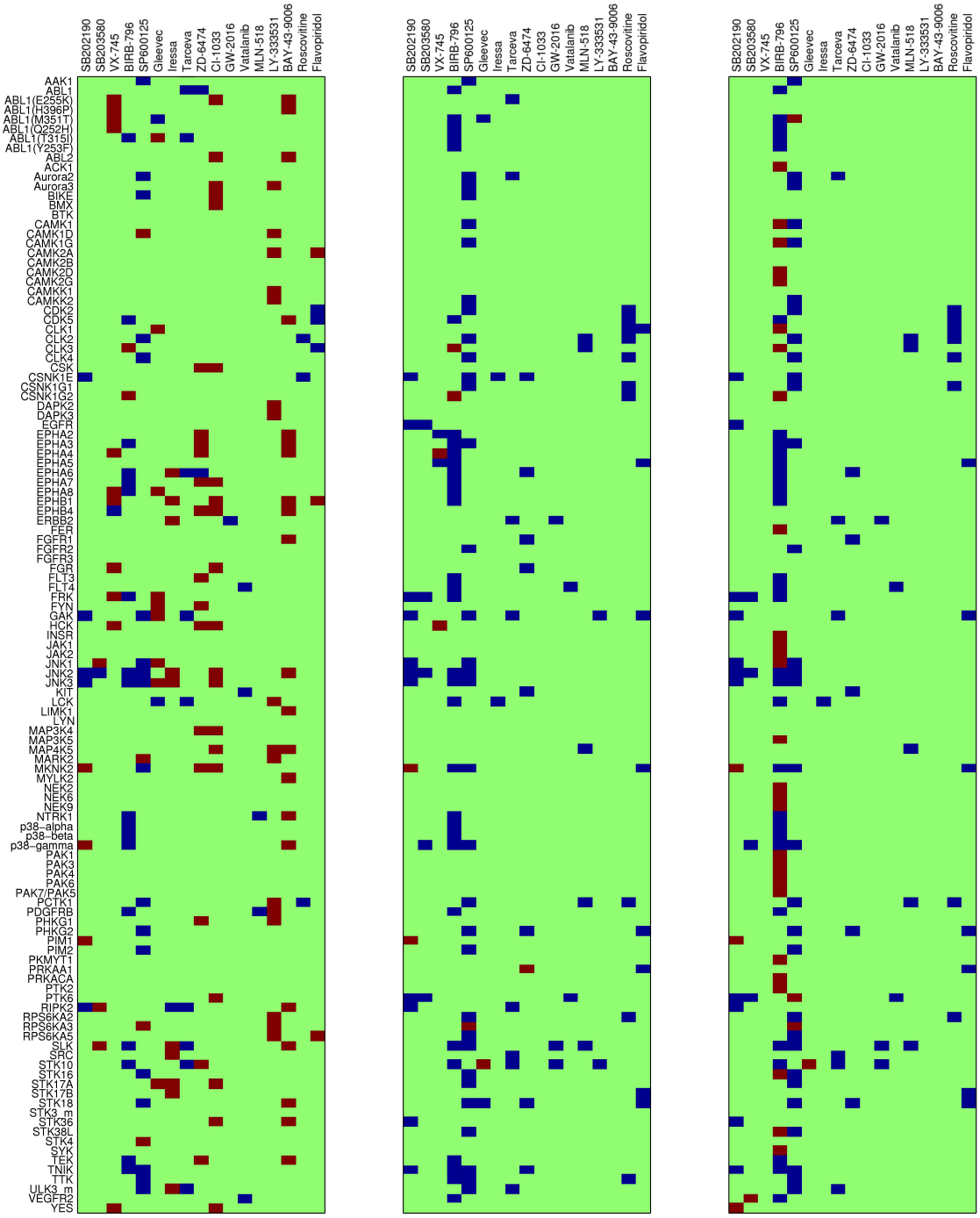

SB202190
SB203580
VX-745
BIRB-796
SP600125
Gleevec
Iressa
Tarceva
ZD-6474
CI-1033
GW-2016
Vatalanib
MLN-518
LY-333531
BAY-43-9006
Roscovitine
Flavopiridol
AAK1
ABL1
ABL1(E255K)
ABL1(H396P)
ABL1(M351T)
ABL1(Q252H)
ABL1(T315I)
ABL1(Y253F)
ABL2
ACK1
Aurora2
Aurora3
BIKE
BMX
BTK
CAMK1
CAMK1D
CAMK1G
CAMK2A
CAMK2B
CAMK2D
CAMK2G
CAMKK1
CAMKK2
CDK2
CDK5
CLK1
CLK2
CLK3
CLK4
CSK
CSNK1E
CSNK1G1
CSNK1G2
DAPK2
DAPK3
EGFR
EPHA2
EPHA3
EPHA4
EPHA5
EPHA6
EPHA7
EPHA8
EPHB1
EPHB4
ERBB2
FER
FGFR1
FGFR2
FGFR3
FGR
FLT3
FLT4
FRK
FYN
GAK
HCK
INSR
JAK1
JAK2
JNK1
JNK2
JNK3
KIT
LCK
LIMK1
LYN
MAP3K4
MAP3K5
MAP4K5
MARK2
MKNK2
MYLK2
NEK2
NEK6
NEK9
NTRK1
p38-alpha
p38-beta
p38-gamma
PAK1
PAK3
PAK4
PAK6
PAK7/PAK5
PCTK1
PDGFRB
PHKG1
PHKG2
PIM1
PIM2
PKMYT1
PRKAA1
PRKACA
PTK2
PTK6
RIPK2
RPS6KA2
RPS6KA3
RPS6KA5
SLK
SRC
STK10
STK16
STK17A
STK17B
STK18
STK3_m
STK36
STK38L
STK4
SYK
TEK
TNIK
TTK
ULK3_m
VEGFR2
YES

Figure 4.5: Matrices of prediction performance, corresponding from left to right to affinity thresholds $K_d = 1\mu M$, $10\mu M$, and $100\mu M$, respectively. The complete affinity profiles of 17 inhibitors independently screened(Fabian *et al.*, 2005) were predicted one by one, using the experimental profiling information on the 17 inhibitors against an 18-kinase subset (different for each inhibitor). Green cells indicate correct predictions; blue, false negatives ("hit" predicted as "no hit"); red, false positives ("no hit" predicted as "hit"). The accuracy percentages shown in the three matrices are 91%, 93% and 93%. Detailed quantitative summary of the accuracy is in Table 4.1.

cuss the practical aspect of the problem: for what kinds of compounds would the predictor work successfully? The 17 compounds we adopted in our analysis represent various types of compounds: SB202190, SB203580, and SP600125 are research compounds; MLN518 is in phase I; VX745, CI-1033, ZD6474, Roscovitine, and Flavopiridol are in phase II; BIRB-796, GW-2016, Vatalanib, LY-333531, and BAY-43-9006 are in phase III; Gleevec, Iressa, and Tarceva are approved drugs(Fabian *et al.*, 2005). This diversity suggests that the predictor would work well in a wide range of compounds. However, there are highly promiscuous compounds, such as Staurosporine, whose profile cannot be fitted into the structure-pharmacology correlation (Section 3.2) and hence our predictor would fail to yield a reliable result.

Another important factor influencing predictor confidence is its robustness or vulnerability to errors introduced in the distance estimation. This is mainly determined by the choice of kinases in the small-scale sample, the affinities for which constitute the sub-profile. Some of the kinases differentiate the inhibitory compounds better than others, i.e. compounds' affinities for them are more rigorous. It is better to include such kinases in the sub-profile, since such basis subsets are less sensitive to the errors in the estimated pharmacological distances. This problem is handled by the optimization process, in which the

subset of kinases that performs best within the training set is chosen (Section 4.1.5).

4.2.2 Validating the predicted affinity profile of a re-designed version of imatinib

To further validate our predictor, we focus on a recently developed kinase inhibitor, WBZ_4 (Figure 4.6), a redesigned version of the powerful anticancer drug imatinib (Gleevec) with higher specificity than the parental compound(Fernández *et al.*, 2007). The prototype compound WBZ_4 does not belong to the drug background used in the high-throughput screening previously adopted as benchmark for our method. The interest in this compound arises because the WBZ_4 design was meant to enhance specificity beyond imatinib levels guided precisely by the same structural markers, the SAHBs that we exploited to calculate pharmacological distances and thus infer cross reactivities. Thus, we now validate our approach by predicting the affinity profile of WBZ_4, and contrasting it with the experimental profile obtained from Ambit's phage-display screening assay reported in ref. of Fernández *et al.* (2007).

The compound WBZ_4 was developed by redesigning imatinib for the purpose of inhibiting the C-Kit kinase, as imatinib does, while avoiding another primary imatinib target, the Bcr-Abl kinase. The latter target has been directly implicated in imatinib's cardiotoxicity(Kerkela *et al.*, 2006). In addition, WBZ_4 was designed to inhibite JNK1, a major target to protect the cardiomyocytes from a mitochondrial collapse induced by Bcr-Abl inhibition(Fernández *et al.*, 2007; Kerkela *et al.*, 2006).

The prototype has been experimentally profiled using the screening methodology based

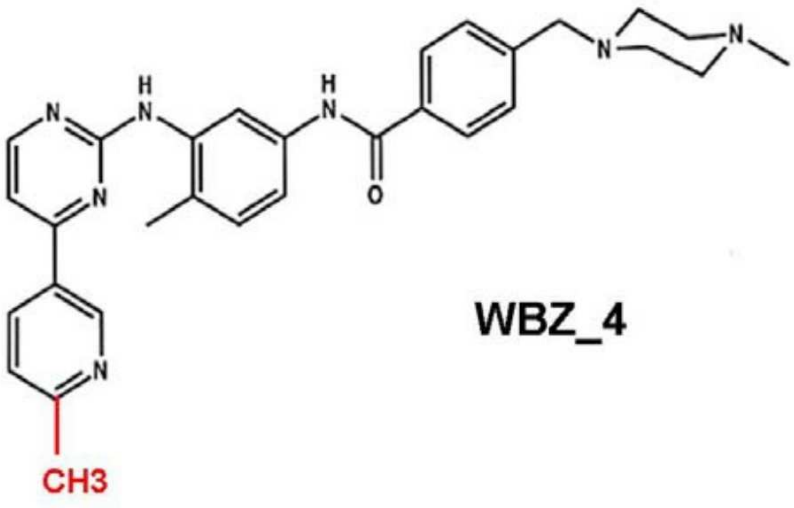


Figure 4.6: Prototype molecule WBZ_4 (N-{5-[4-(4-methyl piperazine methyl)-benzoylamido]-2-methylphenyl}-4-[3-(4-methyl)-pyridyl]-2-pyrimidine amine). It is developed by adding a methyl group (indicated in red) to the imatinib molecule.

on bacteriophage kinase display(Fabian *et al.*, 2005; Fernández *et al.*, 2007). In the prediction, the kinase subset has been optimized in consonance with the drug background of 17 nonpromiscuous compounds extracted from the Ambit's screening (Section4.1.5). The experimental and the predicted results are contrasted in Figure 4.7. The experimental results for the affinities of WBZ_4, reported in ref. Fernández *et al.* (2007), covered 107 of the 119 kinases reported in Fabian *et al.* (2005), excluding ACK1, Aurora2, Aurora3, NTRK1, PRKAA1, PRKACA, STK10, STK18, STK3_m, STK38L, TEK, and ULK3_m. Due to the emphasis in the pharmacological applications of the prototype compound and the clinical significance of achieving nanomolar inhibition, the theoretical predictions were made adopting a stringent threshold $K_{d,threshold} = 100nM$, that is, a hit was recorded as such only if the predicted $K_d < 100nM$. Of the 107 predictions, there is no single false negative and only 2 false positives: LCK and JNK2, as shown in Figure 4.7. This corresponds to an ac-

Figure 4.7: Experimental and predicted results for the affinity profile of WBZ_4 against 107 kinases. The experimental results for the affinities of WBZ_4, reported in Fernández *et al.* (2007), covered 107 of the 119 kinases reported in Fabian *et al.* (2005), excluding ACK1, Aurora2, Aurora3, NTRK1, PRKAA1, PRKACA, STK10, STK18, STK3_m, STK38L, TEK, and ULK3_m. The subset adopted in this prediction has been optimized in advance, within a training set excluding WBZ_4. The optimized subset contains: ABL1(E255K), CAMK1, EPHA8, ERBB2, FLT3, FRK, GAK, INSR, JNK1, KIT, MAP3K4, PDGFRB, PHKG1, PIM1, PRKAA1, RPS6KA2, SLK, SRC. Due to the emphasis in the pharmacological applications of the prototype compound and the clinical significance of achieving nanomolar inhibition, the theoretical predictions were made adopting a stringent threshold $K_{d,threshold} = 100nM$, that is, a hit was recorded as such only if $K_d < 100nM$. Of the 107 predictions, there is no single false negative and only 2 false positives: LCK and JNK2. This corresponds to an accuracy above 98%.

curacy above 98%. Most importantly, our predictor correctly identified C-KIT and JNK1 as primary targets for WBZ_4 and correctly predicted the lack of pharmacological activity against Bcr-Abl, a crucial premise in the planned imatinib redesign geared at curbing its potential cardiotoxicity.

4.2.3 Comparative assessment of performance

One docking-based method was recently reported and claimed to be the best for computing the affinities of inhibitors for homologous receptors (Rockey and Elcock, 2005). The authors compared their predictions with the experimental results published by Fabian *et al.* (2005). and found "a reasonable but not perfect correspondence" (Rockey and Elcock, 2005). Both this work and our profiler predicted the affinity profile of imatinib (Gleevec) against extended lists of kinases, and thus the results can been compared. Benchmarked by the experimental profiles(Fabian *et al.*, 2005), the docking-based prediction contains 9 false negatives (CLK1, CLK4, EPHA8, GAK, JNK1, JNK2, JNK3, STK17A, STK18) and 8 false positives (ACK1, BMX, CSK, FGR, HCK, LYN, RIPK2, YES) out of the 119 kinases[2], while our profiler has only 2 false negatives (ABL1(T315I), STK18) and 1 false positive (STK10) with the $10\mu M$ threshold (Figure 4.5).

Furthermore, the docking-based method is more demanding than our profiler in terms of primary experimental data, since it requires the "high-resolution structure in complex with at least one protein kinase target" for each inhibitor to make appropriate prediction (Rockey and Elcock, 2005). The successful generation of such data is plagued with experi-

[2]The docking-based prediction calculated the affinities of Gleevec for 493 human protein kinases, which are almost the whole human kinome. The 9 false negatives and 8 false positives are only the ones within the 119 kinases tested in the experiments of Fabian *et al.* (2005).

mental uncertainty as crystallization remains a serendipitous craft rather than an automated methodology. By contrast, our profiler requires the affinity sub-profile of the inhibitor against ~ 20 protein kinases, which can be routinely obtained through phage displayers of kinase batteries or other screening methods.

4.3 Conclusion

In this chapter, we introduced a method to predict affinity profiles of inhibitor compounds against entire batteries of human kinases based on a structural descriptor of the targets. The method is rooted in a molecular marker governing drug specificity and promiscuity established in our previous work. A feature-similarity matrix constructed based on the molecular marker is defined across kinase targets and used as an information propagator of a sub-profile where drugs are screened against a small group of kinases. The method reported makes use of distance-geometry techniques, and boils down to determining vectors (kinase profiles) from distances between vectors (feature-similarity distances regarded as surrogates for pharmacological distances between kinases). To guarantee the uniqueness of the solution, some vector coordinates need to be fixed and adopted as constraints. These constraints represent the linear algebra counterpart of the sub-profile. Our *in silico* method enables us to screen large libraries of compounds predicting their profile. To provide the input data, only a limited experimental screening against reduced subset of kinases needs to be performed in advance. Thus, our predictor becomes a valuable tool for lead discovery.

The linear-algebra operations subsumed in our predictor are based on the construction of the informational propagator and hence do not entail any source of errors. The only

systematic source of uncertainty arises from the estimation of pharmacological parameters from structure-based attributes (environmental distances). Given the high accuracy (for a qualitative prediction) of the pharmacological distance estimation, our profiling method should be deemed highly reliable: the accuracy is up to 93% with an optimized choice of kinase subset as starting point, as shown above.

Alternative *in silico* screening methods rooted in docking algorithms are unlikely to match this level of accuracy, not only because of their inherent parametric uncertainty and time expense, but also because kinase binding entails extensive induced fit of the loopy regions within the ATP-pocket (Huse and Kuriyan, 2002). Even those docking algorithms that incorporate induced fits into affinity calculations cannot handle the lengthy loopy regions of kinases, which undergo extensive structural adaptation (Mizutani and Itai, 2004; Mizutani *et al.*, 2006). The induced fit problem as it stands today remains intractable from first-principle aproaches. This is the main reason why we adopted an information-based algorithm for our predictor.

Chapter 5

Redesigning kinase inhibitors to enhance specificity

In this chapter we introduce a technique that turns promiscuous kinase inhibitors into safer drugs. This technique adopted the structrual marker identified in Chapter 3 that is governing the specificity in molecularly targeted drug therapy. This technique is developed besides the recently burgeoning interest in multi-target drugs to treat complex diseases and malignancies. Thus in this chapter we first briefly discuss the assessment of the therapeutic value of promiscuity: Although drug efficacy might not correlate with specificity, it would be risky to welcome promiscuous compounds without a rational strategy to control therapeutic impact. This is the motivation for us to survey approaches to control the therapeutic impact of cross-reactive kinase inhibitors and advocate the application of the selectivity filter by illustrating its cleaning efficacy.

5.1 Assessment of the therapeutic value of promiscuity

Small-molecule inhibitors of protein function are the most common and efficient agents for molecularly targeted therapy geared at treating human disease and malignancy (Dancey and Sausville, 2003; Levitzki and Gazit, 1995; Tibes *et al.*, 2005; Gibbs and Oliff, 1994; Donato and Talpaz, 2000). The undesirable side effects arising from drug cross-reactivity and from the diversity of roles of the target in different biological scenarios prompted researchers to advocate for a "magic-bullet" paradigm (Roth *et al.*, 2004), epitomized by compounds with high binding specificity.

However, there is no obvious correlation between drug specificity and therapeutic index[1](Roth *et al.*, 2004; Frantz, 2005; Keith *et al.*, 2005; Mencher and Wang, 2005; McGovern *et al.*, 2003; Feng *et al.*, 2005). This is especially true for kinase inhibitors since kinases play different signal-transduction roles in different cellular contexts and a favorable inhibition in one scenario may prove fatal in another one (Force *et al.*, 2007). For instance, a specific inhibitor of the Abelson (Abl) kinase would be presumed to be most efficacious in treating chronic myeloid leukemia (CML), since an aberrantly deregulated Abl kinase is a recognized primary target for treating this malignancy (Donato and Talpaz, 2000). However, a systems biology assessment (Force *et al.*, 2007) has recently revealed that Abl inhibition is a culprit for cardiotoxicity. Abl inhibition initiates a signaling cascade that promotes mitochondrial depolarization and hence a health-threatening ATP-depletion in cardiomyocites. Thus, a more cross reactive drug, i. e., one that also inhibits another kinase along the mitochondrial-depolarization pathway (hence blocking it) is expected to

[1]Lethal dose (LD50) over therapeutically effective dose (ED50)

have a higher therapeutic index in treating CML: Higher therapeutic doses may be tolerable due to the removal of cardiotoxic side effects, as recently demonstrated (Fernández *et al.*, 2007; Demetri, 2007; Crunkhorn, 2008).

Furthermore, much effort is recently directed at reassessing the therapeutic value of promiscuity. This paradigm shift is in part motivated by telling cases. For instance, the schizophrenia drug Clozaril (clozapine) works because of its multi-target action, in spite of unpleasant side effects (Roth *et al.*, 2004). Other illustrations of the clinical relevance of cross-reactive ("dirty") drugs have arisen in anticancer therapy: multi-target kinase inhibitors such as Sutent (sunitinib) or Nexavar (sorafenib) have recently received FDA approval (Frantz, 2005; Force *et al.*, 2007), albeit with important caveats[2].

In general, the possibility of exploiting promiscuity is under scrutiny in novel approaches to treat complex disorders such as cancer, depression and cardiovascular disease (Frantz, 2005). Modulating multiple targets simultaneously is often required to alter a clinical phenotype, as biological redundancies and alternative pathways can often bypass the inhibition of a single target or of multiple targets along a single pathway(Hopkins *et al.*, 2006). Thus, a "magic-shotgun" compound targeting multiple proteins may in some instances possess a higher therapeutic index than a specific drug (Roth *et al.*, 2004).

Another argument for promiscuity arises with the observation that dirty or cross-reactive drugs may be more resilient against drug-resistant mutations (Hampton, 2004). Cross reactivity arises because such drugs typically make ligand-target interactions with evolutionarily conserved residues and with backbone groups, and weaker interactions with primary mutated residues (Hampton, 2004; Hopkins *et al.*, 2004). Obviously, targeting conserved

[2]see: http://www.fda.gov/cder/cancer/druglistframe.htm

residues or unspecific parts of the chain (backbone) begets promiscuity.

The safety of dirty drugs, especially dirty kinase inhibitors, is often associated with the possibility of assessing the full extent of their cross reactivity (Owens, 2006). In turn, this assessment is facilitated by the advent of novel high-throughput screening assays such as: a) the kinase assay using the T7-phage expression panels from Ambit Biosciences (Fabian *et al.*, 2005; Karaman *et al.*, 2008); b) a thermal stability shift assay using a 60-Ser/Thr kinase panel (Fedorov *et al.*, 2007a); c) the BioPrint database by Cerep (Krejsa *et al.*, 2003); and d) the living-cell assays of pathway inhibition that assess the impact of the drug on the protein-recruiting capability of the target (MacDonald *et al.*, 2006).

In vitro assays of cross reactivity are of course affected by the complexities of tissue distribution and, generally, by pharmacodynamic issues (Brunton *et al.*, 2005). Yet, the affinity of a drug for a target, being governed by thermodynamics of ligand binding, represents a telling parameter independent of *in vivo* heterogeneities, except for allosteric antagonism (Brunton *et al.*, 2005). On the other hand, crowding, membrane adsorption and other effects can modulate *in vivo* drug concentrations, increasing local levels in different spatial locations. Thus, certain cross-reactivities undetectable *in vitro* may surface in an *in vivo* context, causing unexpected side effects (Brunton *et al.*, 2005; Rishton, 2005). Conversely, tissue and subcellular distribution may prevent a ligand from binding an in-vitro-established target, introducing another caveat in the interpretation of high-throughput screening results.

In spite of timely efforts to establish a paradigm shift, promiscuous drugs lacking controlled specificity obviously carry the burden of life-threatening side effects to a larger extent than their more specific counterparts. Even the most successful anticancer drug

Gleevec (imatinib), with a moderately reduced gamut of primary targets (limited to 5 kinases: Bcr-Abl, C-Kit, Lck, PDGFR, and CSF1R (Fabian *et al.*, 2005; Karaman *et al.*, 2008)), has been recently shown to be potentially cardiotoxic (Force *et al.*, 2007; Hampton, 2004), and labeled as such by the FDA[3]. Not surprisingly, the more promiscuous anticancer kinase inhibitors sunitinib and sorafenib have been also found to be cardiotoxic, even to a larger extent than imatinib (Force *et al.*, 2007; Hampton, 2004). In Ambit screenings (Fabian *et al.*, 2005) sunitinib was shown to bind 79 kinases out of 119 assayed, while sorafenib binds to 41.

If multiple roles of a targeted protein in different cellular contexts may be responsible for side effects (Force *et al.*, 2007), it is only natural that promiscuous compounds would introduce a more uncertain clinical outcome. Hence, it becomes forbiddingly risky to welcome promiscuous compounds into the therapeutic arena without a rational strategy to control their specificity and therapeutic index. This control, in turn, requires the identification of selectivity filters in target space, which should serve as guidance to rational design (Fernández *et al.*, 2007; Fedorov *et al.*, 2007b; Bogoyevitch and Fairlie, 2007; Crespo and Fernández, 2007). In Chapter 3 we identified Solvent-Accessible Hydrogen Bond (dehydron) as one of such filters. Now in this chapter, we advocate for this type of control of therapeutic impact to clean dirty drugs following an integral assessment of the diverse functional roles of targeted proteins. By focusing on kinase targets, we justify this position, offers an avenue to clean dirty kinase inhibitors and demonstrates the feasibility of the proposed approach by reviewing a proof of principle.

[3] see: http://www.fda.gov/medwatch/safety/2006/Gleevec_DHCP_10-19-2006.htm.

5.2 Cleaning cross-reactive drugs by exploiting selectivity filters

Most protein drug targets have paralogs, that is, proteins that share a common ancestor with the target and have diverged away from it after speciation (Chen *et al.*, 2007). Thus, kinases, the widespread cancer targets, belong to common-ancestry groups (families) which typically share the same fold and basic structural features. This structural conservation often results in unexpected cross reactivities that may lead to undesired side effects (Fabian *et al.*, 2005; Karaman *et al.*, 2008; Griffin, 2005).

In principle, much of the cross reactivity may be removed by drug redesign guided by the identification of structural features that promote promiscuity and nonconserved features that enable paralog discrimination. This approach has been attempted with promising results (Fernández *et al.*, 2007; Demetri, 2007; Crunkhorn, 2008), and supports our proposal for "cleaning" kinase inhibitors.

There have been a number of paralog-discriminating nonconserved features that may be exploited as selectivity filters, that is, as targetable differences (Fedorov *et al.*, 2007b; Bogoyevitch and Fairlie, 2007). One approach uses high-resolution crystal structures of kinases in complex with non-ATP ligands to identify unique structural motifs in the purported targets. For example, a unique helical insert has been found in the activation loop of the NEK2 and MPSK1 kinases following the conserved DFG catalytic triad (Fedorov *et al.*, 2007b).

Specific inhibitors that target inactive kinase conformations have also been developed targeting the "DFG out" conformation (Liu and Gray, 2006). In this unique conformation,

the position of the phenylalanine (F) residue, located at the start of the activation loop, is flipped with respect to the active conformation, so that it points inwards within the ATP pocket. This mode of association is illustrated by the binding of imatinib to the inactive Abl kinase (Noble *et al.*, 2004). In addition to the DFG-out motif, nonconserved structural features within the inactive ensemble should perhaps be exploited to achieve paralog specificity. The inactive conformations of kinases make them more discernible, while the active conformation reveals fewer discriminatory features since it is constrained to be catalytically functional and hence more conserved. While targeting the inactive conformation may be a logical choice, there are also advantages in targeting the active conformation. The latter requires structure conservation for functional purposes, and hence it is less tolerant to drug-resistant mutations (Noble *et al.*, 2004). The substrate-discriminatory amino acid variations that tell kinases apart are mostly located in loopy regions framing the ATP-pocket, rather than in the pocket itself, making them less accessible targets (Crespo and Fernández, 2007; Noble *et al.*, 2004).

In this regard, another way of approaching the specificity problem is the design of allosteric kinase inhibitors (Bogoyevitch and Fairlie, 2007). These ligands are typically more specific than ATP-competitive inhibitors, since they bind to residues outside the ATP-pocket, which are typically less conserved (Bogoyevitch and Fairlie, 2007).

In this chapter, we present a strategy adopting the molecular basis for specificity introduced in Chapter 3, Solvent-Accessible Hydrogen Bond (dehydron). The distribution of dehydrons in the structure may be turned into an operational selectivity filter for two reasons:

- Dehydrons may be targeted by drugs that further wrap them by binding to the protein

(c.f. Section 2.2.2).

- Dehydrons are not conserved across paralogs (c.f. Section 3.2).

To assess conservation, we align the paralog structures and examine the microenvironments of the aligned hydrogen bonds. Typically, while the bonds are conserved across paralogs, their packing is not, and hence there are differences in the dehydron distribution (Chen *et al.*, 2007; Fernández and Scheraga, 2003). The structure conservation across proteins within kinase families enables the alignment (Chen *et al.*, 2007).

Therefore, the differences in dehydron distribution on the protein surfaces constitute promising selectivity filters to clean dirty inhibitors through redesign. Thus, their stickiness arises since the association of a suitable "wrapping" ligand to the target protein containing the dehydron entails further removal of water surrounding the latter (Chen *et al.*, 2007).

5.3 A first validation of the approach

While various selectivity filters have been identified for the druggable kinome (Fedorov *et al.*, 2007b; Bogoyevitch and Fairlie, 2007), only the nonconserved wrapping patterns were adopted to re-design drugs in order to re-focus their impact on clinical targets (Crespo and Fernández, 2007). For example, imatinib has been re-designed to curb its potential cardiotoxicity (Force *et al.*, 2007; Kerkela *et al.*, 2006) by introducing a wrapping modification that removes its inhibitory impact on one of its primary targets, the Bcr-Abl kinase (Fernández *et al.*, 2007; Demetri, 2007; Crunkhorn, 2008). The inhibition of this kinase in cardiomyocites has been shown to be causative of imatinib's cardiotoxicity (Kerkela *et al.*, 2006). On the other hand, imatinib is known to be particularly effective in treating chronic

myeloid leukemia (CML) precisely through its inhibitory impact on the Bcr-Abl kinase, a constitutively active aberrant chimera (Schindler *et al.*, 2000; Gambacorti-Passerini *et al.*, 1997). Furthermore, imatinib has additional primary targets (C-Kit, Lck, PDGFR, and CSF1R (Fabian *et al.*, 2005; Karaman *et al.*, 2008)), with at least one, C-Kit, of proven clinical relevance (Demetri, 2007). This information inspired imatinib redesign with one goal: avoiding the Bcr-Abl kinase while retaining activity towards C-Kit, the primary target to treat GIST (gastrointestinal stromal tumors) (Demetri, 2007). The structural alignment of Abl (PDB.1FPU) and C-Kit (PDB.1T46) complexed with imatinib reveals a nonconserved dehydron C673-G676 in C-Kit which aligns with a well-wrapped M318-G321 hydrogen bond in Abl (Fernández *et al.*, 2007) (Figure 5.1). This difference in wrapping at the catalytic loop prompted the development of a methylated variant of imatinib that hampers Abl inhibition while re-focusing the impact on c-Kit. The molecular basis for target discrimination was established *in vitro* and through *in vivo* assays for antitumor activity. The therapeutic impact of the modified compound was confirmed in novel GIST animal models, also corroborating a significant reduction in cardiotoxicity (Fernández *et al.*, 2007; Demetri, 2007; Crunkhorn, 2008).

This proof of principle reveals that dehydrons are indeed targetable features and hence opens up the possibility of exploiting differences in dehydron patterns to guide the cleaning of promiscuous inhibitors through molecular redesign.

We thus suggest a general strategy to clean dirty inhibitors by introducing “wrapping” chemical modifications that preserve the drug chemotype (Hopkins *et al.*, 2006) while targeting unique dehydrons. The proof of principle here described serves as a first validation for this approach.

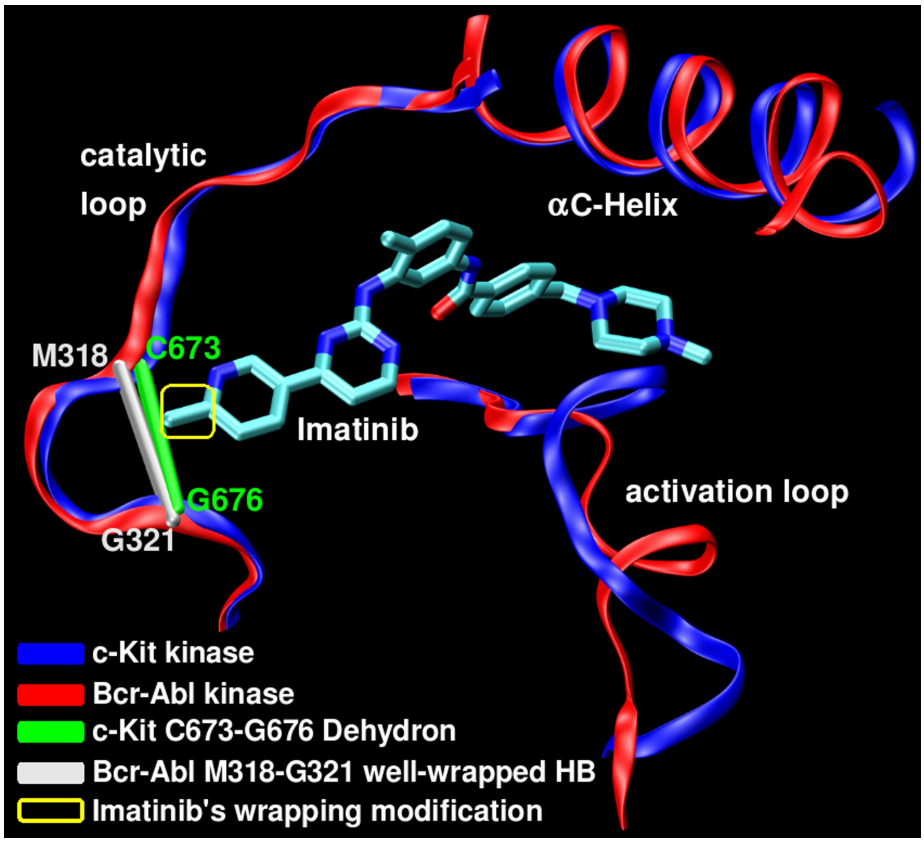


Figure 5.1: Aligned backbones (ribbon representation) of Bcr-Abl (PDB.1FPU, red) and C-Kit (PDB.1T46, blue) kinases in their respective structurally adapted imatinib complexes. The nonconserved dehydron C673-G676 (green) in C-Kit aligns with the well wrapped M318-G321 hydrogen bond (gray) in Abl, and has been targeted by a methylation "wrapping" modification of imatinib (yellow highlight) to achieve specificity.

5.4 A workable approach

Most kinase inhibitors are in principle susceptible of being turned into selective wrappers of packing defects through minor chemical modification and without altering their chemotype. Thus, clinically relevant compounds with considerable cross reactivities such as sunitinib (55% hits over number of proteins screened (Karaman *et al.*, 2008)), dasatinib

(30%), EKB-569 (20%), sorafenib (20%) or erlotinib (15%) may be redesigned into drugs with enhanced specificity using the wrapping design concept. As an illustration, let us focus on cleaning EKB-569.

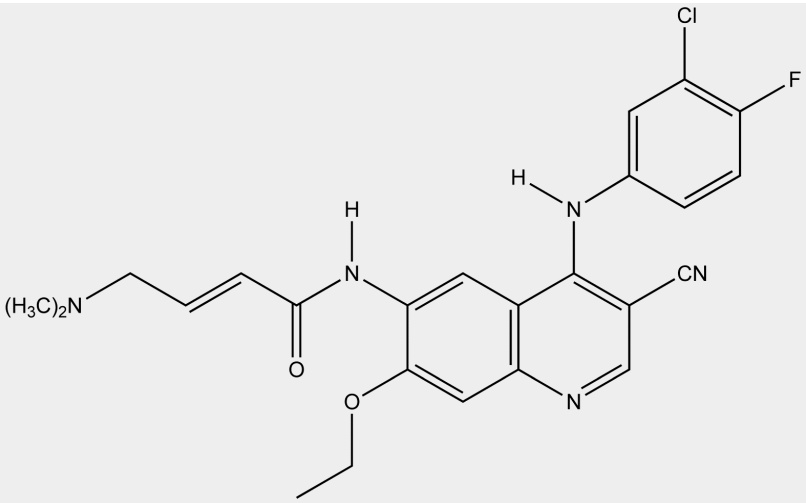


Figure 5.2: Kinase inhibitor EKB-569 (Wyeth-Ayerst, (Torrance *et al.*, 2000)), a major inhibitor of the epidermal growth factor receptor (EGFR) kinase

The irreversible kinase inhibitor EKB-569 (Wyeth-Ayerst, (Torrance *et al.*, 2000)) was launched as a major inhibitor of the epidermal growth factor receptor (EGFR) kinase (IC_{50}=38.5nM). Thus, its therapeutic interest to treat non-small cell lung cancer (NSCLC), colorectal neoplasia and other EGFR-dependent solid tumors became apparent[4]. Phase I and II trials for such therapeutic applications are currently closed (Torrance *et al.*, 2000; Erlichman *et al.*, 2006). Recent high-throughput screening using a battery of 119 T7-phage expressed kinases (Fabian *et al.*, 2005) revealed 25 sub-micromolar targets for EKB-569, making it a promiscuous drug with likely side effects. Other compounds such as Iressa (gefitinib) and Tarceva (erlotinib) share the same "4-anilinoquinoline" chemotype (Wiss-

[4]see: http://www.cancer.gov/search/ResultsClinicalTrialsAdvanced.aspx?protocolsearchid=4056579

ner *et al.*, 2003), yet they are more specific EGFR inhibitors (Fabian *et al.*, 2005). The latter two gained FDA approval as anti-NSCLC agents[5].

Here we applied to EKB-569 the general strategy as follows.

1. Identify and verify the "dirty part", i.e., the source for the promiscuity.

2. Clean up the "dirty part" by removing the structure features in the compound that cause promiscuity.

3. Search for a targetable feature (dehydron in this work) in the intended target kinase that is not conserved across the kinase paralogs.

4. Further modify the compound by introducing "wrapping" group that preserves the drug chemotype while targets the unique dehydron chosen in the previous step.

In the following sections, we will explicitly explain the cleaning procedure step by step.

5.4.1 Identification of promiscuity source

The EKB-569 promiscuity can be traced to its intermolecular interactions with highly conserved residues within the EGFR kinase family. As shown in Figure 5.3, the terminal acryl group in the ligand plays the role of electrophile in the irreversible Michael-adduct reaction with the nucleophile-conserved residues Cys/Ser in the EGFR-paralog kinases. The water-solublizing terminal N-dimethyl group of EKB-569 may also accelerate such addition, serving as an intramolecular base catalyst for Michael reaction with the Cys or

[5]see: http://www.fda.gov/cder/cancer/druglistframe.htm

Ser residues, due to the spatial proximity(Wissner *et al.*, 2003). Another source of EKB-569 promiscuity is the intermolecular electrostatic interaction between its cyanide group and the gatekeeper residue (Thr or Met), typically conserved within the family (Figure 5.3).

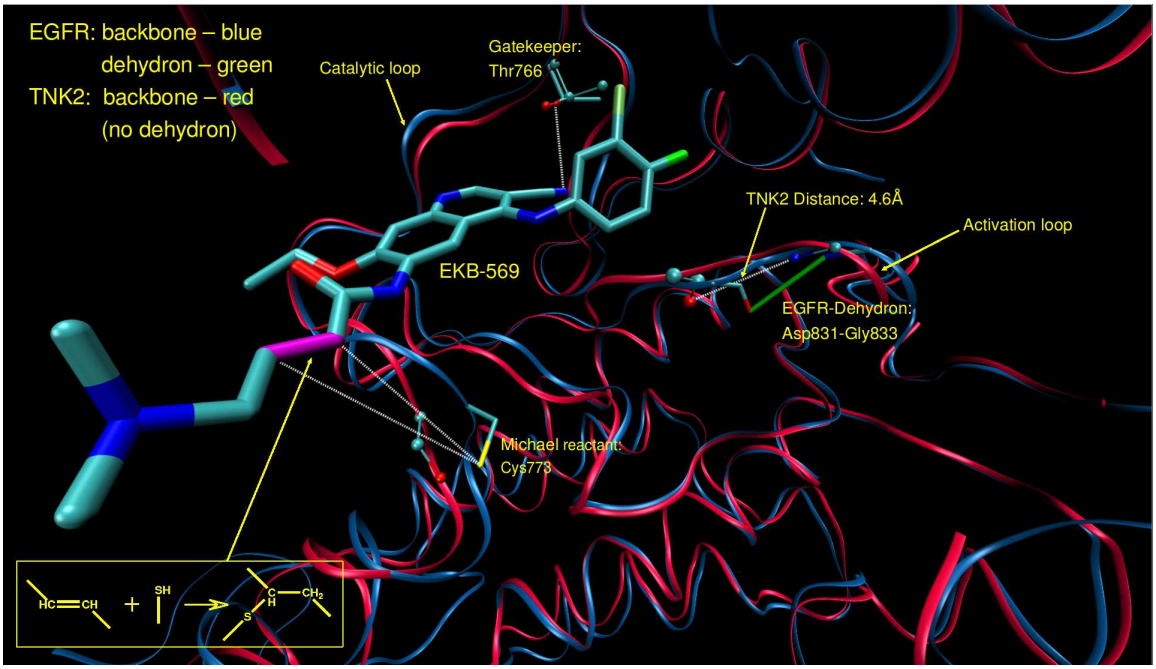


Figure 5.3: Structural alignment of EGFR kinase (blue ribbon representation, atoms in licorice) and the paralog TNK2 kinase (red ribbon representation, atoms in balls and sticks), complexed with EKB-569 (licorice). Atoms are depicted following standard color convention (chlorine in green, fluorine in light green). One source of EKB-569 promiscuity is the terminal acryl group (magenta), the electrophile group involved in the Michael reaction with the nucleophile-conserved residues Cys/Ser in EGFR and its paralog kinases. The other source of drug promiscuity is the intermolecular electrostatic interaction between its cyanide group and the conserved gatekeeper residue (Thr/Met) in the target protein. The wrapping pattern of EGFR includes the poorly conserved Asp831-Gly833 dehydron that may be targeted to achieve selectivity. TNK2 contains the same two promiscuity-fostering features, while lacking the dehydron at the locus where EGFR contains the specificity-promoting feature. Thus, targeting the latter will ensure a discriminatory binding of EGFR without hitting TNK2, as experimentally corroborated.

5.4.2 Statistical verification of promiscuity source

To validate these observations, we assessed the correlation between the affinities of EKB-569 for the 48 paralogs of EGFR reported in PDB and the extent of residue conservation at the Michael reaction site and at the gatekeeper position. To do so, we aligned each paralog structure with the EGFR structure and examined residues that align with Cys773 (Michael reactant) and Thr766 (gatekeeper). The aligning residues are listed in Table 5.1. Then we built a logistic regression model (Agresti, 1996) to assess the correlation between the affinities of EKB-569 for the EGFR paralogs and the extent of residue conservation at the two key sites.The logistic regression model is built as follows.

Explanatory Variables: There are two explanatory variables: the types of residues aligning with Cys773 (Michael reactant) and Thr766 (gatekeeper), respectively. Since the Michael reaction can take place only if the residue aligning with Cys773 is a Michael reactant, i.e., Cys or Ser, we let the first explanatory variable, X_1, to be 1 if the residue is Cys or Ser, otherwise to be 0. Similarly, we let the second explanatory variable $X_2 = 1$ if the residue aligning with Thr766 is Thr or Met (possible intermolecular electrostatic interaction with the cyanide group), and $X_2 = 0$ otherwise.

Responding Variables: Naturally, the responding variable should indicate the affinity of EKB-569 towards the 48 EGFR-paralogs. Here the responding variable Y represents the affinity in a "hit/no-hit" level: $Y = 1$ if the dissociation constant (K_d) is smaller than $3\mu M$ according to Fabian *et al.* (2005), and $Y = 0$ otherwise.

Null Hypothesis: The null hypothesis of the model is that there is NO statistically significant correlation between the responding variable and the explanatory variables. This

Table 5.1: Data for the logistic regression model

Kinase	PDB	Michael reaction site[¶]	Gatekeeper position[§]	X_1	X_2	Y
ABL1	1IEP	ASN	**THR**	0	1	1
AURKA	1MUO	THR	LEU	0	0	0
BTK	1K2P	**CYS**	**THR**	1	1	1
CAMK1G	2JAM	GLU	**MET**	0	1	0
CDK2	1AQ1	ASP	PHE	0	0	0
CDK5	1UNG	ASP	PHE	0	0	0
CLK1	1Z57	**SER**	PHE	1	0	1
CLK3	2EU9	ASN	PHE	0	0	1
CSNK1G2	2C47	**SER**	LEU	1	0	0
DAPK2	1ZWS	GLU	LEU	0	0	0
DAPK3	2J90	GLU	LEU	0	0	0
EGFR	1M17	**CYS**	**THR**	1	1	1
EPHA2	1MQB	ALA	**THR**	0	1	0
ERBB2	1OVC	**CYS**	**THR**	1	1	1
FGFR1	1AGW	ASN	VAL	0	0	0
FGFR2	1GJO	ASN	VAL	0	0	0
FLT3	1RJB	ASP	PHE	0	0	0
FYN	2DQ7	**SER**	**THR**	1	1	1
HCK	1AD5	**SER**	**THR**	1	1	1
INSR	1GAG	ASP	**MET**	0	1	0
JAK2	2B7A	**SER**	**MET**	1	1	1
JNK1	2NO3	ASN	**MET**	0	1	0
JNK3	1PMN	ASN	**MET**	0	1	0
KIT	1T45	ASP	**THR**	0	1	0
LCK	2OF2	**SER**	**THR**	1	1	1
MAP3K5	2CLQ	**SER**	**MET**	1	1	0
MKNK2	2AC3	**SER**	PHE	1	0	0
NEK2	2JAV	ASP	**MET**	0	1	1
P38-α	1A9U	ASP	**THR**	0	1	0
P38-γ	1CM8	ASP	**MET**	0	1	0
PAK1	1YHV	**SER**	**MET**	1	1	1
PAK4	2CDZ	ALA	**MET**	0	1	0
PAK6	2C30	ALA	**MET**	0	1	0
PAK7/PAK5	2F57	ALA	**MET**	0	1	0
PDGFRb	1LWT	ASP	**THR**	0	1	
PIM1	1XQZ	ASP	LEU	0	0	0
PIM2	2IWI	ASP	LEU	0	0	0
PKAC-α	2GU8	GLU	**MET**	0	1	
PTK2	2ETM	GLU	**MET**	0	1	0
RPS6KA5	1VZO	GLU	LEU	0	0	0
SLK	2J51	ALA	ILE	0	0	1
SRC	2SRC	**SER**	**THR**	1	1	1
STK10	2J7T	ALA	ILE	0	0	1
STK16	2BUJ	THR	LEU	0	0	0
SYK	1XBC	PRO	**MET**	0	1	1
TIE2(TEK)	1FVR	ASN	ILE	0	0	1
TNK2(ACK)	1U46	**SER**	**THR**	1	1	1
VEGFR2	2P2H	ASN	VAL	0	0	0

hypothesis will be checked by the p-value of the model.

The date for the logistic regression model is listed in Table 5.1.

We fit the model with the data in Table5.1 using a web-based logistic regression model fitting tool [6] and revealed that the EKB-569 affinity is indeed dictated by these two sources of promiscuity: p-value=0.0015 indicates that we can reject the null hypothesis at a confidence level as high as 99.75%. The goodness of the fitting result is listed in Table 5.2. Thus, the hypothesis that the terminal acryl group and the cyanide group in EKB-569 (Figure 5.3) are indeed the "dirty" moieties responsible for promiscuity is statistically confirmed.

Table 5.2: Goodness of the Logistic Regression Model Fitting

Chi Square	degree of freedom	p-value
13.0285	2	0.0015

5.4.3 Clean-up of promiscuity source

Having identified the sources of promiscuity, we proceeded to clean EKB-569 by introducing the following chemical modifications (Figure 5.4):

a) Replace the double bond (the Michael acceptor) in the acryl group with a single bond.

b) Replace the cyanide group with a methyl.

Since one hydrophilic group (the cyanide) is replaced by a hydrophobic group (the methyl) in step (b), it may decrease the solubility of the compound. To compensate against the insolubility due to the hydrophobic effect of the methyl group, we make an additional revision to the molecular structure:

[6] http://statpages.org/logistic.html

c) Shorten the hydrophobic tail at the No.7 position of the nitroquinoline such that the tail in the new compound is only a methoxy.

After the clean-up step, the compound structure is as follows:

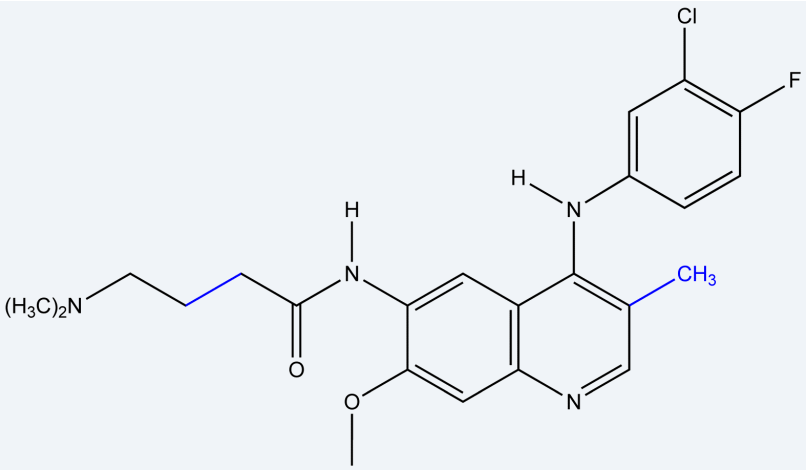


Figure 5.4: EKB-569 with its promiscuity sources removed (the replacing parts are colored in blue)

5.4.4 Choosing unique dehydron and introducing "wrapping" modification

After removing the sources of promiscuity, we then seek wrapping-based targetable feature in the original target of EKB-569, EGFR kinase. With the targetable feature identified, the next step is to introduce a wrapping modification in the drug to target the intended feature. When EGFR is crystallized in the induced-fit conformation generated by an inhibitor (erlotinib) that shares EKB-569's 4-anilinoquinoline chemotype (PDB.1M17), we now find only four accessible dehydrons within the binding pocket (see Figure 5.5):

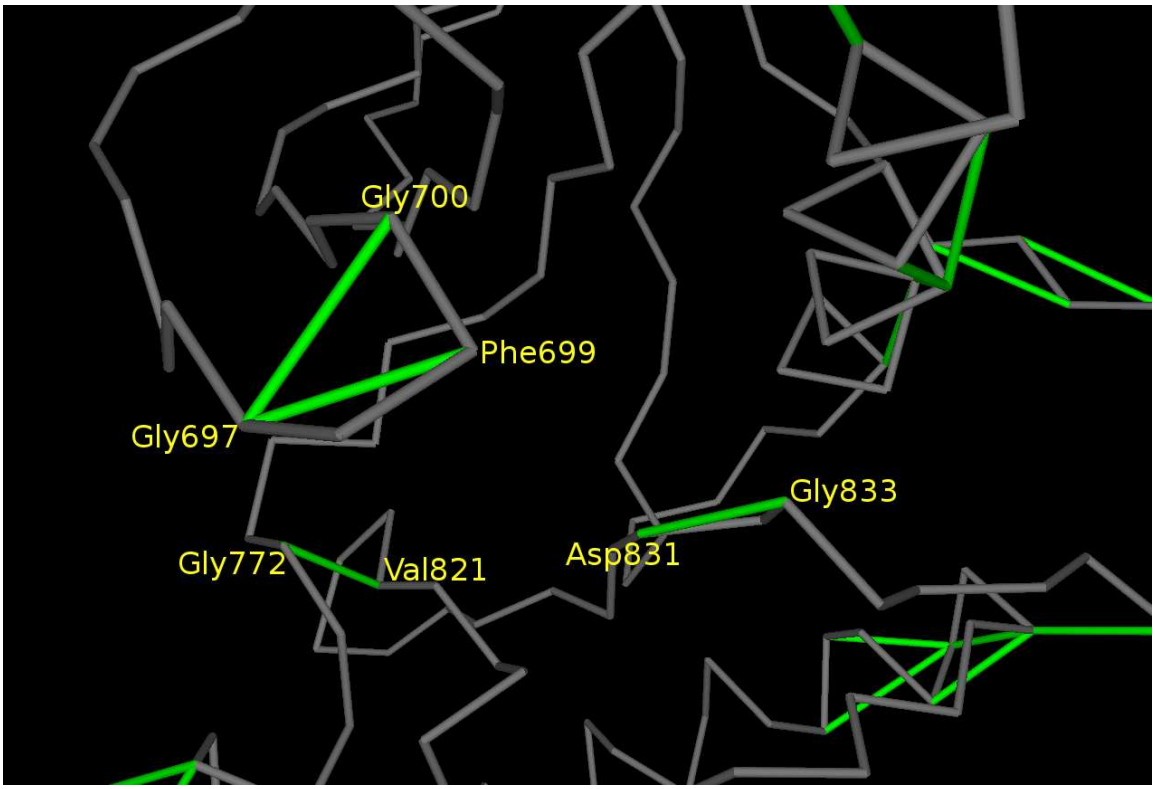


Figure 5.5: Accessible dehydrons within the binding pocket of EGFR (1M17.pdb). Only the backbone is illustrated (in gray) for clarity. Dehydrons are indicated by green virtual bond connecting the α-carbons. There are other dehydrons present in the EGFR kinase, but only those accessible ones within the binding pocket are labeled.

Asp831-Gly833, Gly697-Phe699, Gly697-Gly700, Gly772-Val821.

By examining the conservation of these four dehydrons across the 48 EGFR-paralogs, we found the least conserved is dehydron Asp831-Gly833. Only 11 paralogs retain this dehydron:

AURKA, CLK3, EGFR, ERBB2, FYN, LCK, PAK6, PAK7/PAK5, PIM2, SLK, STK10.

Figure 5.3 shows an example kinase TNK2, which does not retain the dehydron. Based on the dehydron-conservation analysis, we choose this dehydron as the nonconserved se-

lectivity feature to be targeted. To do so, we appended a methyl group at position 3 of the terminal benzene ring that acts as a wrapper or protector of such feature (Figure 5.6 and 5.7).

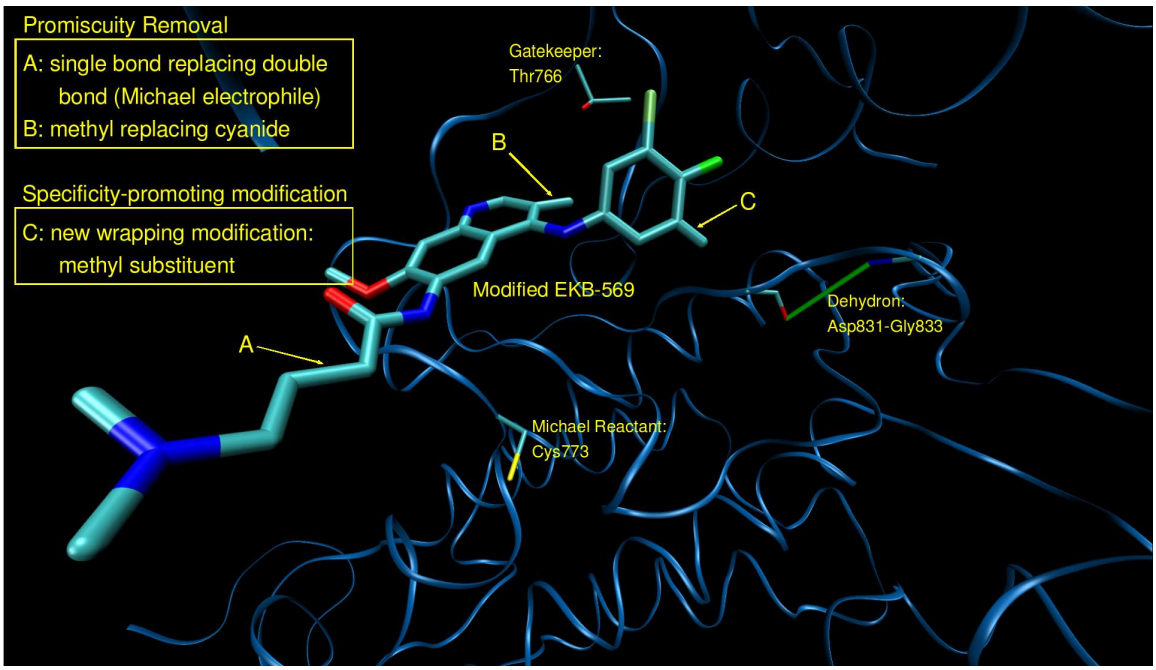


Figure 5.6: Structural features promoting selectivity in EGFR kinase guiding EKB-569 cleaning redesign. EGFR kinase structure (same representation as above) complexed with the prototype EKB-569 re-designed inhibitor (licorice representation). To remove EKB-569 promiscuity, the acrylic double bond (Michael electrophile) is replaced by a single bond and the gatekeeper-interacting cyanide is replaced by a methyl. To selectively target EGFR, a methyl group is added to the terminal benzene ring as a wrapper of the barely conserved Asp831-Gly833 dehydron.

5.4.5 Synthesis of the redesigned EKB-569

This re-designed EKB-569 molecule was synthesized in Eli Lilly and Company following a pathway that recapitulates the EKB-569 synthesis (Wissner *et al.*, 2003). The

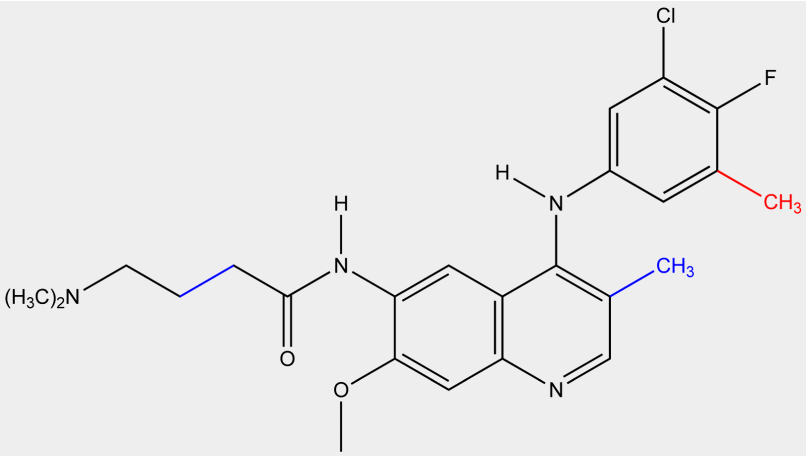
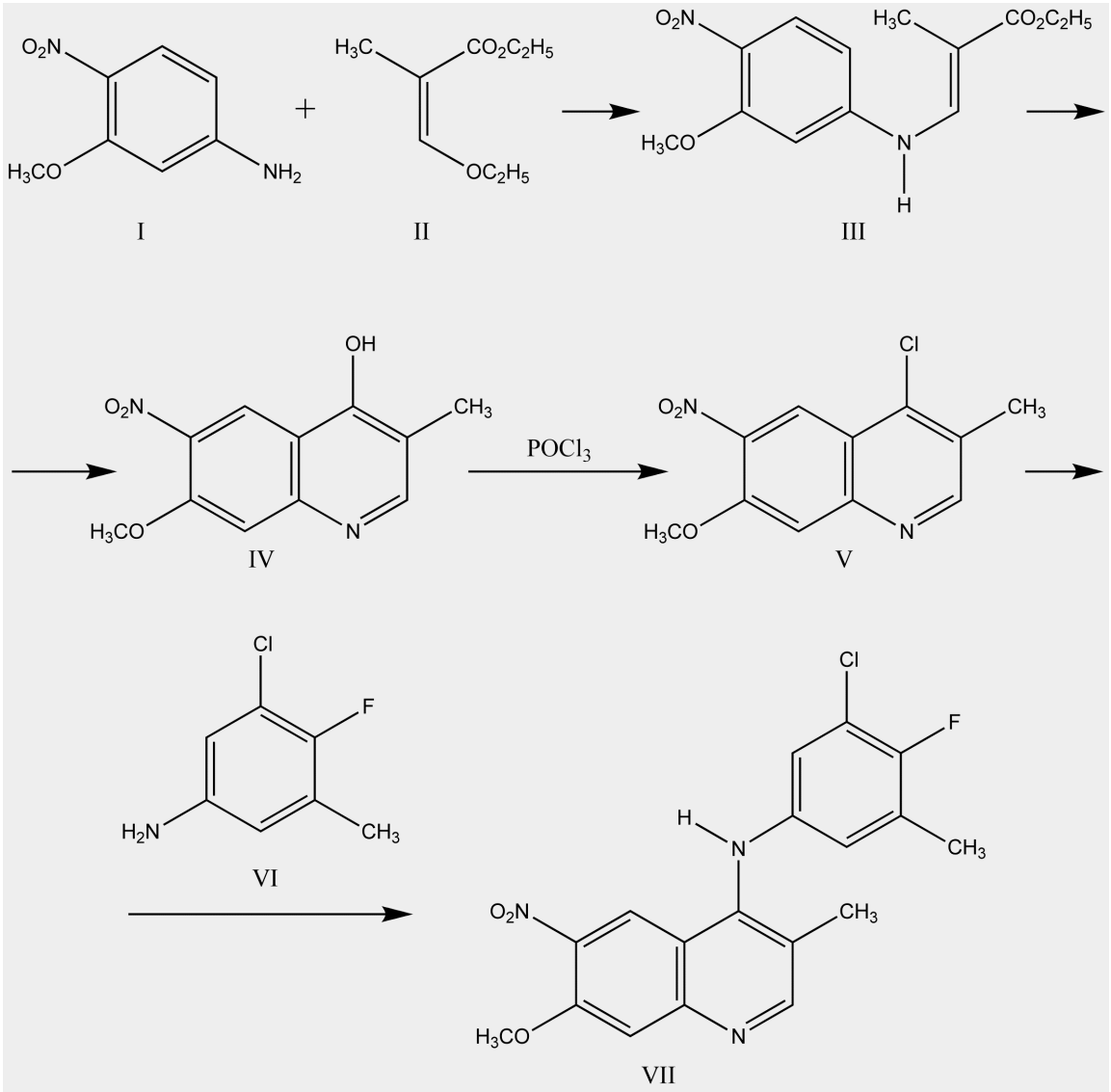


Figure 5.7: EKB-569 with its promiscuity sources removed (the replacing parts are colored in blue) and a methyl group (in red) added as a wrapper of the dehydron Asp831-Gly833 (see Figure 5.6) in EGFR.

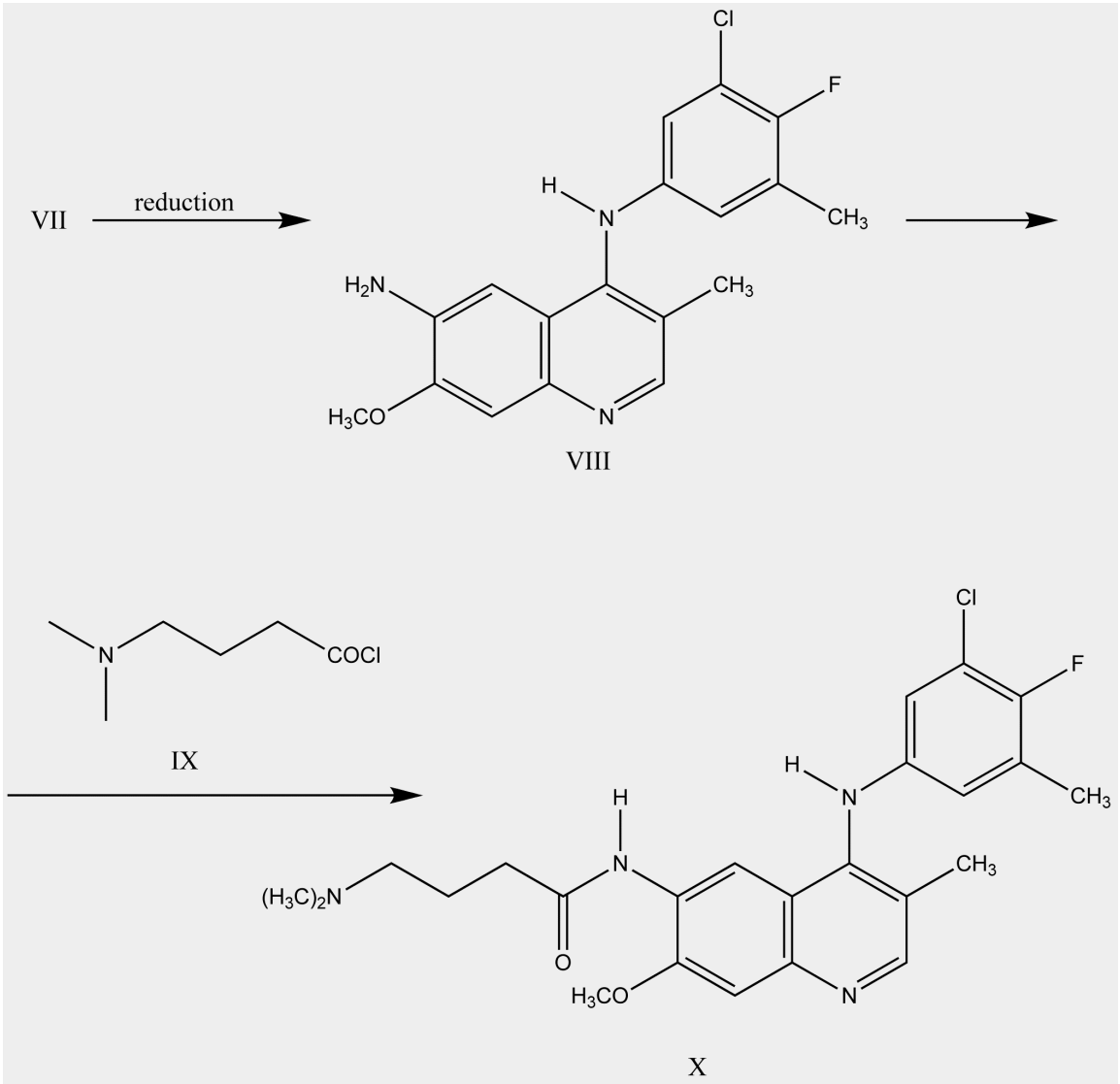
synthesis is schemed in Figure 5.8 and the details is as follows.

7-methoxy-4-hydroxy-6-nitroquinoline-3-methyl (IV)

A mixture 54.4 g (0.324 mol) of I and 72.1 g (0.456 mol) of II in 210 mL of toluene was refluxed for about 16 hs. The reaction was cooled in an ice bath, and the product was filtered. This was washed with three portions of ether and then dried to give 85.4 g of intermediate compound III (94.1%) as a mixture of cis/trans isomers which could be recrystallized, in 80% yield, from 2-methoxyethanol. A portion of this compound (34.4 g, 0.123 mol) was added as a solid to 2.5 L of refluxing (256 °C) dowtherm under N_2 in a 5 L three-necked flask equipped with a mechanical stirrer and a thermometer under nitrogen. The reaction mixture was stirred vigorously at this temperature for 1.25 hs and then cooled to room temperature. The thick reaction mixture was diluted with 2 L of ether, filtered, and washed with ether to yield 21.5 g (74.7% from III) of IV.



I: 4-nitro-3-methoxyaniline
II: ethyl(ethoxymethylene)methylacetate
IV: 7-methoxy-4-hydroxy-6-nitroquinoline-3-methyl
VI: 3-methyl-4-fluoro-5-chloro-aniline
VII: 4-(3-methyl-4-fluoro-5-chlorophenylamino)-7-methoxy-6-nitroquinoline-3-methyl



IX: 4-dimethylamino-butanoyl-chloride
X: 4-dimethylamino-butanoic-acid-[4-(5-chloro-4-fluoro-3-methylphenylamino)-3-methyl-7-methoxy-quinoline-6-yl]-amide

Figure 5.8: Synthetic pathway of the redesigned EKB-569 inhibitor.

7-methoxy-4-chloro-6-nitroquinoline-3-methyl (V)

18 g (77 mmol) of IV was refluxed with 120 mL of $POCl_3$ under N_2 for 2.5 hs in a 1 L round-bottomed flask. TLC (EtOAc: hexane, 1:1) showed no starting material remaining. The excess $POCl_3$ was removed by rotary evaporation. The flask containing the solid residue was cooled in an ice bath, and 600 mL of CH_2Cl_2 was added to dissolve the residue. The resulting cold solution was added into a vigorously stirred solution of 250 mL ice-cold saturated K_2CO_3 and stirred for 30 min. The organic layer was separated, washed, dried ($MgSO_4$), and evaporated to give 17.3 g (88.8%) of V.

4-(3-methyl-4-fluoro-5-chlorophenylamino)-7-methoxy-6-nitroquinoline-3-methyl (VII)

A solution of 24.4 g (96.5 mmol) of V and 15.2 g (96.5 mmol) of VI in 900 mL of 2-propanol was refluxed under N2 for 3.5 hs. TLC (EtOAc: hexane, 1:1) showed no starting material remaining. After standing at room-temperature overnight, the solid was collected by filtration and washed with 2-propanol and ether to give 37.2 g (93.8%) of VII as HCl salt.

4-(3-methyl-4-fluoro-5-chlorophenylamino)-7-methoxy-6-aminoquinoline-3-methyl (VIII)

The hydrochloride VII (34.2 g, 91.2 mmol) was mixed with 35.7 g (638 mmol) of iron powder. A solution of 43.9 g (820 mmol) of NH_4Cl in 280 mL of water was added followed by 985 mL of CH_3OH. The mixture was refluxed with mechanical stirring under N_2 for 4 hs at which time TLC indicated complete reduction. The reaction mixture was filtered hot, and solids were washed with 500 mL of boiling CH_3OH. After the combined filtrate was evaporated, the residue was partitioned between 1.5 L of warm ethyl acetate and 700 mL of

saturated sodium bicarbonate solution. The organic layer was dried over $MgSO_4$, treated with activated charcoal, filtered, and evaporated to give a residue which was recrystallized from $CHCl_3$-hexanes giving 28.6 g (90.9%) of VIII.

4-dimethylamino-butanoic-acid-[4-(5-chloro-4-fluoro-3-methylphenylamino)-3-methyl-7-methoxy-quinoline-6-yl]-amide (X)

A solution of 18.9 g (54.9 mmol) of VIII and 11.5 mL (65.9 mmol) of N,N-diisopropyl ethylamine in 366 mL of anhydrous THF was stirred under N_2 in an ice bath as a solution of 12.1 g (79.6 mmol) of the acid chloride (IX) in 183 mL of THF was added over 15 min. The reaction vessel was sealed and stored in the freezer overnight. The solution was evaporated, and the residue was partitioned between saturated $NaHCO_3$ and EtOAc. The organic layer was separated, washed, dried ($MgSO_4$), and passed through a thin layer of silica gel. The obtained solid was refluxed with 400 mL of CH_3OH for 0.5 h. After being cooled to room temperature, the solid was collected and washed with CH_3OH followed by hexane to give 21.1 g (77.8%) of X as HCl salt: Melting point 197-200; MS (ES+) m/z 459; 1H-NMR (DMSO-d6) δ: 9.73 (bs, 1H), 9.61 (s, 1H), 9.00 (s, 1H), 8.54 (s, 1H), 7.41 (m, 3H), 7.24 (m, 1H), 6.83 (dt, J = 5.6, J = 15.4 Hz, 1H), 4.35 (s, 3H), 3.07 (d, J = 5.1 Hz, 2H), 1.41 (s, 3H), 1.31 (s, 3H); Anal. (C24H28N4O2ClF HCl) C, H, N.

5.4.6 Prediction of the redesigned EKB-569's profile

Based on the selectivity filter provided by the wrapping patterns, we predict the affinity profile of our prototype. The prediction is based on the conservation of the EGFR dehydron Asp831-Gly833 wrapped by the prototype (but not by EKB-569) and the existence

of steric hindrances with nonpolar groups in the catalytic loop of the target structures. We predicted as "hits" only those targets with a conserved dehydron in such position and no steric hindrance. Of the 13 kinases retaining the dehydron, only 7 hits are predicted:

CLK3, EGFR, ERBB2, FYN, LCK, SLK, STK10.

The other six all have steric hindrance with the compound. In fact, these three kinases do not bind with the original EKB-569 either (Fabian *et al.*, 2005). In the cases where the residues aligning with EGFR's Asp831-Gly833 are not engaged in a dehydron or a well-wrapped hydrogen bond, we examined whether such dehydron can be induced upon ligand binding with minimal structural adaptation. This happens only if two residues are not forming a dehydron or a hydrogen bond in the pdb file but would form a dehydron (not a well-wrapped hydrogen bond) when the loop makes small changes which change the distance between or the relative orientation of the two residues. We found that only in BTK, PTK2 and SYK kinases this dehydron can be induced upon drug binding with no steric hindrance with the catalytic loop, so they represent other "possible" hits. Thus, we predicted 7 "hits" and 3 "possible hits". The details of the prediction is listed in Table 5.3

5.4.7 Experimentally screening the redesigned EKB-569's profile

To validate our strategy, a comparative high-throughput screening was conducted at $10\mu M$ EKB-569 and prototype over a battery of 228 human kinases displayed in a T7-bacteriophage-expressing library (Ambit Bioscience, San Diego, CA). Figure 5.9 shows the complete experimental screening result.

The experimentally obtained affinity profile for the prototype agrees almost perfectly with our predicted profile (see the last two columns of Table 5.3): There is only one false

Table 5.3: Prediction and Experimental Validation

Kinase	PDB	Wrapping classification	Steric hindrance	Predicted affinity	Experimental affinity	Match prediction-experiment
ABL1	2GQG	-		NO HIT	NO HIT	YES
AURKA	1MQ4	Dehydron	YES	NO HIT	NO HIT	YES
BTK	1K2P	Possibly induced	NO	Possible HIT	HIT	YES
CAMK1G	2JAM	-		NO HIT	NO HIT	YES
CDK2	1AQ1	-		NO HIT	NO HIT	YES
CDK5	1UNG	-		NO HIT	NO HIT	YES
CLK1	1Z57	-		NO HIT	NO HIT	YES
CLK3	2EU9	Dehydron	NO	HIT	HIT	YES
CNSK1G2	2C47	-		NO HIT	NO HIT	YES
DAPK2	2A2A	-		NO HIT	NO HIT	YES
DAPK3	2J90	-		NO HIT	NO HIT	YES
EGFR	1M17	Dehydron	NO	HIT	HIT	YES
EPHA2	1MQB	-		NO HIT	NO HIT	YES
ERBB2	1OVC	Dehydron	NO	HIT	HIT	YES
FGFR1	1AGW	-		NO HIT	NO HIT	YES
FGFR2	1GJO	-		NO HIT	NO HIT	YES
FLT3	1RJB	Possibly induced	YES	NO HIT	NO HIT	YES
FYN	2DQ7	Dehydron	NO	HIT	NO HIT	NO
HCK	1QCF	-		NO HIT	NO HIT	YES
INSR	1GAG	-		NO HIT	NO HIT	YES
JAK2	2B7A	-		NO HIT	NO HIT	YES
JNK1	1UKH	-		NO HIT	NO HIT	YES
JNK3	1PMN	-		NO HIT	NO HIT	YES
KIT	1PKG	-		NO HIT	NO HIT	YES
LCK	1QPC	Dehydron	NO	HIT	HIT	YES
MAP3K5	2CLQ	-		NO HIT	NO HIT	YES
MKNK2	2AC3	-		NO HIT	NO HIT	YES
NEK2	2JAV	-		NO HIT	NO HIT	YES
P38-α	1DI9	-		NO HIT	NO HIT	YES
P38-γ	1CM8	-		NO HIT	NO HIT	YES
PAK1	1YHV	-		NO HIT	NO HIT	YES
PAK4	2CDZ	-		NO HIT	NO HIT	YES
PAK6	2C30	Dehydron	YES	NO HIT	NO HIT	YES
PAK7/PAK5	2F57	Dehydron	YES	NO HIT	NO HIT	YES
PDGFRB	1LWP	-		NO HIT	NO HIT	YES
PIM1	1YXT	-		NO HIT	NO HIT	YES
PIM2	2IWI	Dehydron	YES	NO HIT	NO HIT	YES
PKAC-α	2GU8	-		NO HIT	NO HIT	YES
PTK2	2ETM	Possibly induced	NO	Possible HIT	HIT	YES
RPS6KA5	1VZO	-		NO HIT	NO HIT	YES
SLK	2J51	Dehydron	NO	HIT	HIT	YES
SRC	2SRC	-		NO HIT	NO HIT	YES
STK10	2J7T	Dehydron	NO	HIT	HIT	YES
STK16	2BUJ	Possibly induced	YES	NO HIT	Not screened	
SYK	1XBB	Possibly induced	NO	Possible HIT	NO HIT	NO
TIE2	1FVR	-		NO HIT	NO HIT	YES
TNK2	1U46	-		NO HIT	NO HIT	YES
VEGFR2	2P2H	-		NO HIT	NO HIT	YES

positive (FYN), and one semi-false negative: SYK, which is predicted as "possible hit" but not bound by the redesigned EKB-569.

The confirmation of the predictions by the experiments suggests that we have not only successfully cleaned the dirty inhibitor EKB-569, but also controlled the therapeutic impact in a predictable manner.

Summary of this section: We successfully cleaned the dirty inhibitor EKB-569 using an approach based on a selectivity filter. This was accomplished by first removing the chemical features that promote promiscuity. Subsequently, we introduced a wrapping modification to target a non-conserved dehydron in the intended target and made the prototype more selective than the parental compound.

5.5 Conclusion

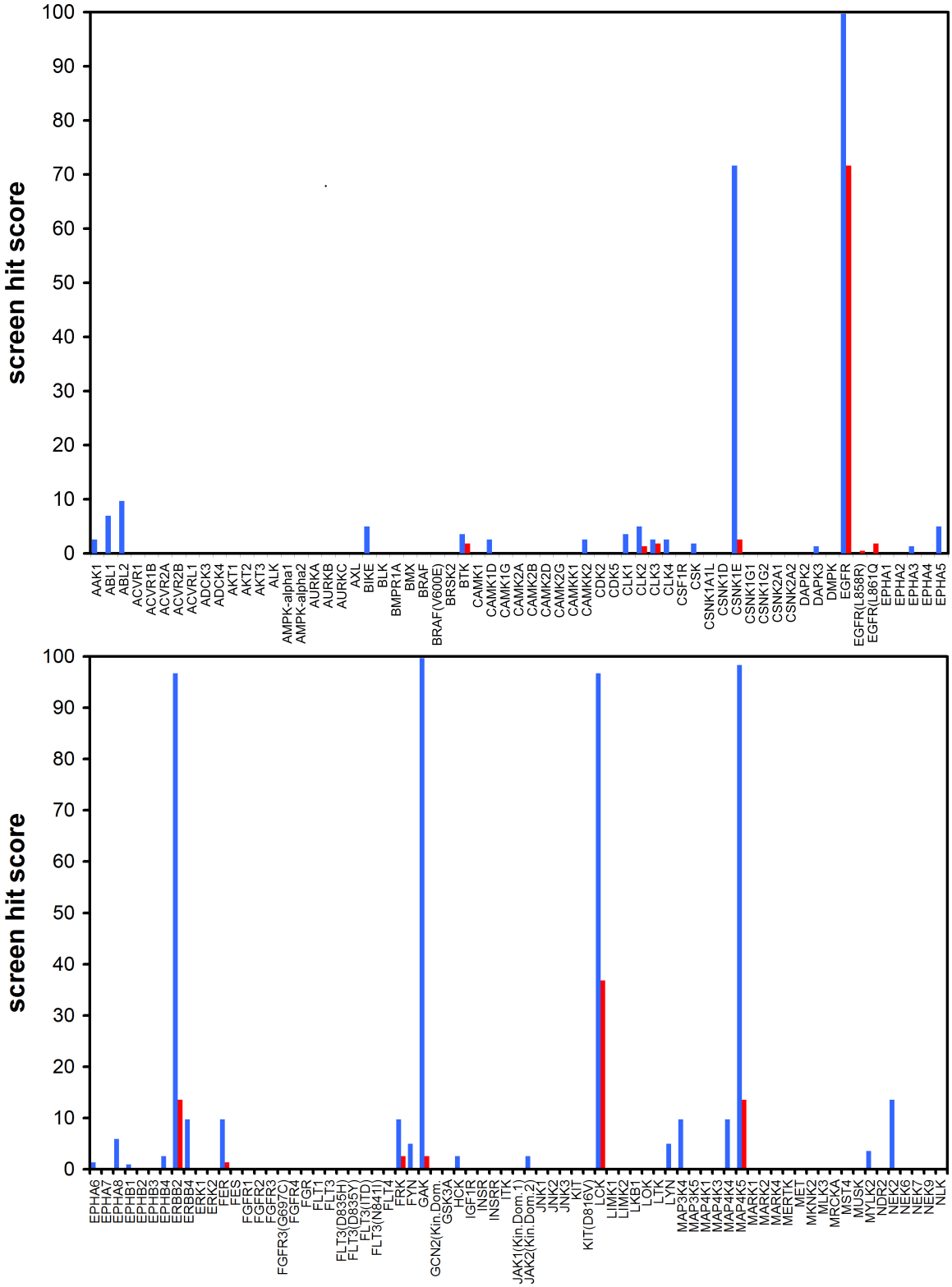
As we have argued, therapeutic efficacy may not correlate with drug specificity, as revealed in treatments of complex disorders and malignancies (Roth *et al.*, 2004; Frantz, 2005; Keith *et al.*, 2005; Mencher and Wang, 2005; McGovern *et al.*, 2003; Feng *et al.*, 2005; Hopkins *et al.*, 2006). This has motivated a reassessment of the therapeutic value of promiscuity and may well trigger a paradigm shift, from "magic bullets" to multi-target therapies (Frantz, 2005; Hopkins *et al.*, 2006). These conceptual leaps are supported by the advent of novel high-throughput screening technologies enabling an assessment of cross reactivities (Fabian *et al.*, 2005; Karaman *et al.*, 2008; Fedorov *et al.*, 2007a; Krejsa *et al.*, 2003), of the breadth of therapeutic impact and hidden phenotypes (MacDonald *et al.*, 2006), and of likely side effects (Force *et al.*, 2007; Kerkela *et al.*, 2006).

However, knowing the realm of activity of these "magic shotguns" (Roth *et al.*, 2004) is insufficient to warrant broad therapeutic application. The uncertainties they introduce in treatment outcome are likely broader than the more limited side effects associated with more specific drugs (Force *et al.*, 2007; Kerkela *et al.*, 2006). Not surprisingly, the pharmaceutical industry still holds a parochial approach in the face of such safety uncertainties (Frantz, 2005; Hopkins *et al.*, 2006).

At this juncture, multi-target molecular therapies could only be welcomed if their target selectivity can be controlled to curb side effects and treatment uncertainties. Thus, there is a niche in emerging biotechnologies for novel approaches to clean promiscuous drugs in order to achieve a tighter specificity control. We have shown that such approaches are in principle feasible by redesigning dirty drugs guided by novel selectivity filters (Fernández *et al.*, 2007; Demetri, 2007; Crunkhorn, 2008). Future developments will no doubt lead to more effective ways of confining cross reactivity to targets of clinical relevance as better structural markers for specificity are discovered. The drug-redesign exercise and the proof of principle described here suggest that cleaning a dirty drug guided by basic new concepts is in principle possible and will hopefully inspire further efforts in this regard.

Promiscuity might become a welcomed feature in kinase-inhibitor design but only provided that cross reactivity can be held under tight control through the implementation of selectivity filters. We advocate that at least one such filter can be exploited rationally to enhance the specificity of promiscuous compounds, such as sunitinib, dasatinib, EKB-569, sorafenib and erlotinib, towards their respective primary targets: KIT/VEGFR2, ABL/SRC, EGFR, VEGFR2 and EGFR. This approach requires information subtler than a structural characterization because it takes into account packing differences across targets.

The dearth of structural information on available targets does not constitute a hindrance to the applicability of this approach because packing information can be inferred directly from sequence, by taking into account the fact that poorly wrapped regions of the structure have a propensity to be disordered (Fernández and Berry, 2004; Pietrosemoli and Crespo, 2007).


screen hit score
100
90
80
70
60
50
40
30
20
10
0
AAK1
ABL1
ABL2
ACVR1
ACVR1B
ACVR2A
ACVR2B
ACVRL1
ADCK3
ADCK4
AKT1
AKT2
AKT3
ALK
AMPK-alpha1
AMPK-alpha2
AURKA
AURKB
AURKC
AXL
BIKE
BLK
BMPR1A
BMX
BRAF
BRAF(V600E)
BRSK2
BTK
CAMK1
CAMK1D
CAMK1G
CAMK2A
CAMK2B
CAMK2D
CAMK2G
CAMKK1
CAMKK2
CDK2
CDK5
CLK1
CLK2
CLK3
CLK4
CSF1R
CSK
CSNK1A1L
CSNK1D
CSNK1E
CSNK1G1
CSNK1G2
CSNK2A1
CSNK2A2
DAPK2
DAPK3
DMPK
EGFR
EGFR(L858R)
EGFR(L861Q)
EPHA1
EPHA2
EPHA3
EPHA4
EPHA5
screen hit score
100
90
80
70
60
50
40
30
20
10
0
EPHA6
EPHA7
EPHA8
EPHB1
EPHB2
EPHB3
EPHB4
ERBB2
ERBB4
ERK1
ERK2
FER
FES
FGFR1
FGFR2
FGFR3
FGFR3(G697C)
FGFR4
FGR
FLT1
FLT3
FLT3(D835H)
FLT3(D835Y)
FLT3(ITD)
FLT3(N841I)
FLT4
FRK
FYN
GAK
GCN2(Kin.Dom.
GSK3A
HCK
IGF1R
INSR
INSRR
ITK
JAK1(Kin.Dom.2)
JAK2(Kin.Dom.2)
JNK1
JNK2
JNK3
KIT
KIT(D816V)
LCK
LIMK1
LIMK2
LKB1
LOK
LTK
LYN
MAP3K4
MAP3K5
MAP4K1
MAP4K3
MAP4K4
MAP4K5
MARK1
MARK2
MARK4
MERTK
MET
MKNK2
MLK3
MRCKA
MST4
MUSK
MYLK2
NDR2
NEK2
NEK6
NEK7
NEK9
NLK

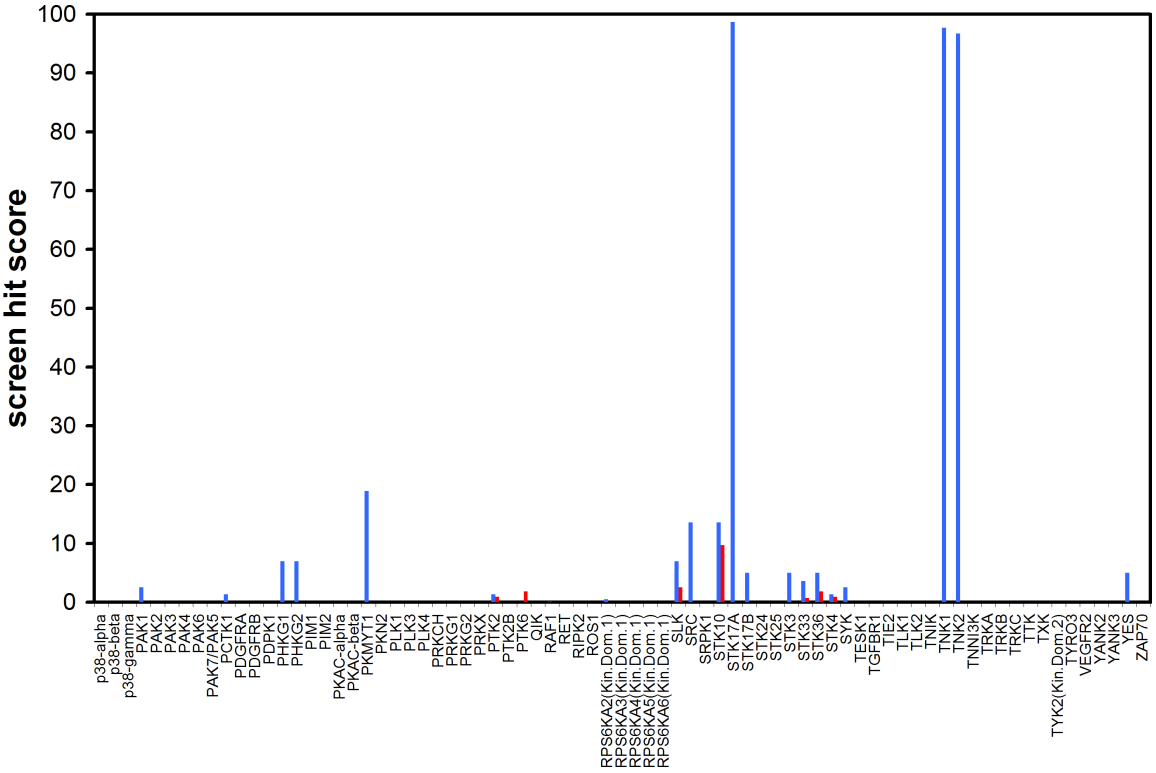


Figure 5.9: Affinity profile of the original/redesigned EKB-569 inhibitors. High-throughput screening at 10 μM of original EKB-569 (blue) and redesigned EKB-569 (red) over a battery of 228 human kinases displayed in a T7-bacteriophage-expressing library (Ambit Bioscience, San Diego, CA). The screening assay of EKB-569 (blue) was used as control. Hit values are reported as percentage bound kinase.

Bibliography

A. Agresti. *An Introduction to Categorical Data Analysis*, chapter 4: Logistic Regression. Wiley Interscience, 1996.

J. Bain, M. McLauchlan, H. Eliott, and P. Cohen. The specificities of protein kinase inhibitors: an update. *Biochem. J.*, 371:299–204, 2003.

J. M. Berg, J. L. Tymoczko, and L. Stryer. *Biochemistry*. W. H. Freeman and Company, New York, 5th edition, 2002.

K. H. Bleicher, H. J. Böhm, Muller K., and A. I. Alanine. Hit and lead generation: beyond high-throughput screening. *Nat. Rev.*, 2:369–378, 2003.

M. A. Bogoyevitch and D. P. Fairlie. A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding. *Drug Discov. Today*, 12: 622–633, 2007.

R. Bonneau, C. E. Straus, C. A. Rohl, D. Chivian, P. Bradley, L. Malmstrom, T. Robertson, and D. Baker. *De novo* prediction of three-dimensional structures for major protein families. *J. Mol. Biol.*, 322:65–78, 2002.

C. Braken, L. M. Iakoucheva, P. R. Romero, and A. K. Dunker. Combining prediction, computation and experiment for the characterization of protein disorder. *Curr. Op. Str. Biol.*, 14:570–576, 2004.

L. L. Brunton, J. S. Lazo, and K. L. Parker. *Goodman & Gilman's The Pharmacological Basis of Therapeutics*. McGraw-Hill Companies, 11th edition, 2005.

J. Chen, X. Zhang, and A. Fernández. Molecular basis for promiscuity and specificity in the druggable kinome. *Bioinformatics*, 23:563–572, 2007.

D. Chivian, D. E. Kim, L. Malmstrom, J. Schonbrun, C. A. Rohl, and D. Baker. Prediction of CASP6 structures using automated Robetta protocols. *Proteins*, 61:157–166, 2005.

C. Chothia. Hydrophobic bonding and accessible surface area in proteins. *Nature*, 248: 338–339, 1974.

A. Crespo and A. Fernández. Kinase packing defects as drug targets. *Drug Discov. Today*, 12:917–923, 2007.

S. Crunkhorn. Anticancer Drugs: Redesigning kinase inhibitors. *Nat. Rev. Drug Disc.*, 7: 120–121, 2008.

J. Dancey and E. A. Sausville. Issues and progress with protein kinase inhibitors for cancer treatment. *Nat. ReV. Drug DiscoVery*, 2:296–313, 2003.

G. D. Demetri. Structural reengineering of imatinib to decrease cardiac risk in cancer therapy. *J. Clin. Invest.*, 117:3650–3653, 2007.

G. D. Demetri, M. V. Mehren, C. D. Blanke, A. D. Van den Abbeele, B. Eisenberg, P. J. Roberts, M. C. Heinrich, D. A. Tuveson, S. Singer, M. Janicek, J. A. Fletcher, S. G. Silverman, S. L. Silberman, R. Capdeville, B. Kiese, B. Peng, S. Dimitrijevic, B. J. Druker, C. Corless, C. D. M. Fletcher, and H. Joensuu. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. *N. Engl. J. Med.*, 347:472–480, 2002.

A. B. Dietz, L. Souan, G. J. Knutson, M. R. Bulur, P. A. Litzow, and S. Vuk-Pavlović. Imatinib mesylate inhibits T-cell proliferation in vitro and delayed-type hypersensitivity in vivo. *Blood*, 104:1094–1099, 2004.

N. J. Donato and M. Talpaz. Clinical use of tyrosine kinase inhibitors: Therapy for chronic myelogenous leukemia and other cancers. *Cancer Res.*, 6:2965–66, 2000.

J. Drews. Drug Discovery: A Historical Perspective. *Science*, 287:1960–1964, 2000.

B. J Druker. Molecularly targeted therapy: have the floodgates opened? *Oncologist*, 9: 357–360, 2004.

C. Erlichman, M. Hidalgo, J. P. Boni, P. Martins, S. E. Quinn, C. Zacharchuk, P. Amorusi, A. A. Adjei, and E. K. Rowinsky. Phase I Study of EKB-569, an irreversible inhibitor of the epidermal growth factor receptor, in patients with advanced solid tumors. *J. Clin. Oncol.*, 24:2232–2260, 2006.

H. Escriva, F. Delaunay, and Laudet V. Ligand binding and nuclear receptor revolution. *Bioessays*, 22:357–360, 2000.

D. Fabbro and C. G. Garcia-Echeverria. Targeting protein kinases in cancer therapy. *Curr. Opin. Drug Discovery Dev.*, 5:701–712, 2002.

M. A. Fabian, W. H. Biggs, D. K. Treiber, C. E. Atteridge, M. D. Azimioara, M. G. Benedetti, T. D. Carter, Ciceri P., Edeen P. T., Floyd M., Ford J. M., Galvin M., Gerlach J. L., R. M. Grotzfeld, S. Herrgard, D. E. Insko, M. A. Insko1, Lai A. G., J-M Lelias, S. A. Mehta, Z. V. Milanov, A. M. Velasco, L. M. Wodicka, H. K. Patel, P. P. Zarrinkar, and D. J. Lockhart. A small molecule kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.*, 23:329–336, 2005.

O. Fedorov, B. Marsden, V. Pogacic, P. Rellos, S. Müller, A. N. Bullock, J. Schwaller, M. Sundström, and S. Knapp. A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. *Proc. Natl. Acad. Sci. U. S. A.*, 104:20523–20528, 2007a.

O. Fedorov, M. Sundström, B. Marsden, and S. Knapp. Insights for the development of specific kinase inhibitors by targetd strcutural genomics. *Drug Discov. Today*, 12:365–372, 2007b.

B. Y. Feng, A. Shelat, T. N. Doman, R. K. Guy, and B. K. Shoichet. High throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.*, 1:146–148, 2005.

A. Fernández. Keeping dry and crossing membranes. *Nat. Biotech.*, 22:1081–1084, 2004.

A. Fernández and R. S. Berry. Molecular dimension explored in evolution to promote proteomic complexity. *Proc. Natl. Acad. Sci. USA*, 101:13460–13465, 2004.

A. Fernández and S. Maddipati. A priori inference of cross reactivity for drug-targeted kinases. *J. Med. Chem.*, 49:3092–3100, 2006.

A. Fernández, A. Sanguino, Z. Peng, E. Ozturk, J. Chen, A. Crespo, S. Wulf, A. Shavrin, C. Qin, J. Ma, J. Trent, Y. Lin, H. D. Han, L. S. Mangala, J. A. Bankson, J. Gelovani, A. Samarel, W. Bornmann, A. K. Sood, and G. Lopez-Berestein. An anticancer C-kit kinase inhibitor is re-engineered to make it more active and less cardiotoxic. *Journal of Clinical Investigation*, 117:4044–4054, 2007.

A. Fernández and H. A. Scheraga. Insufficiently dehydrated hydrogen bonds as determinats of protein interactions. *Proc. Natl. Acad. Sci. U. S. A.*, 100:113–118, 2003.

A. Fernández, T. R. Sosnick, and A. Colubri. Dynamics of hydrogen bond desolvation in protein folding. *J. Mol. Biol.*, 321:659–675, 2002.

T. Force, D. S. Krause, and R. A. Van Etten. Molecular mechanisms of cardiotoxicity of tyrosine kinase inhibition. *Nat. Rev. Can.*, 7:332–344, 2007.

R. Fraczkiewicz and W. Braun. Exact and efficient analytical calculation of the accessible surface areas and their gradient for macromolecules. *J. Comput. Chem.*, 19:319–333, 1998.

S. Frantz. Drug discovery: playing dirty. *Nature*, 437:942–943, 2005.

A. Gabriele and G. L. King. Protein kinase C inhibitors in the treatment and prevention of diabetic complications. *Curr. Opin. Endocrinol. Diabetes*, 8:197–204, 2001.

C. Gambacorti-Passerini, P. le Coutre, L. Mologni, M. Fanelli, C. Bertazzoli, E. Marchesi, M. Di Nicola, A. Biondi, G. M. Corneo, D. Belotti, E. Pogliani, and N. B. Lydon. Inhibition of the ABL kinase activity blocks the proliferation of BCR/ABL + leukemic cells and induces apoptosis. *Blood Cells Mol. Dis.*, 23:380–394, 1997.

J. Gibbs and A. Oliff. Pharmaceutical research in molecular oncology. *Cell*, 79:193–198, 1994.

J. D. Griffin. Interaction maps for kinase inhibitors. *Nat. Biotech.*, 23:308–309, 2005.

T. Hampton. "Promiscuous" anticancer drugs that hit multiple targets may thwart resistance. *JAMA*, 292:419 – 422, 2004.

D. G. Higgins, J. D. Thompson, and T. J Gibson. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.*, 266:383–402, 1996.

C. W. V. Hogue. Cn3D: a new generation of three-dimensional molecular structure viewer. *Trends Biochem. Sci.*, 22:314–316, 1997.

A. L. Hopkins, J. S. Mason, and J. P Overington. Can we rationally design promiscuous drugs? *Curr. Op. Struct. Biol.*, 16:127–136., 2006.

A. L. Hopkins, J. Ren, J. Milton, R. J. Hazen, J. H. Chan, D. I. Stuart, and D. K. Stammers. Design of non-nucleoside inhibitors of HIV-1 reverse transcriptase with improved drug resistance properties. *J. Med. Chem.*, 47:5912–5922, 2004.

M. Huse and J. Kuriyan. The conformational plasticity of protein kinases. *Cell*, 109: 275–282, 2002.

M. W. Karaman, S. Herrgard, D. K. Treiber, P Gallant, C. E. Atteridge, B. T. Campbell, K. W. Chan, P. Ciceri, M. I. Davis, P. T. Edeen, R. Faraoni, M. Floyd, J. P. Hunt, D. J. Lockhart, Z. V. Milanov, M. J. Morrison, G. Pallares, H. K. Patel, S. Pritchard, L. M.

Wodicka, and P. P. Zarrinkar. A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotech.*, 26:127–132, 2008.

C. T. Keith, A. A. Borisy, and B. R. Stockwell. Multicomponent therapeutics for networked systems. *Nat. Rev. Drug. Discov.*, 4:71–78, 2005.

R. Kerkela, L. Grazette, R. Yacobi, C. Iliescu, R. Patten, C. Beahm, B. Walters, S. Shevtsov, S. Pesant, F. J. Clubb, A. Rosenzweig, R. N. Salomon, R. A. Van Etten, J. Alroy, J.-B. Durand, and T. Force. Cardiotoxicity of the cancer therapeutic agent imatinib mesylate. *Nat. Med.*, 12:908–916, 2006.

D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev.*, 3:935–949, 2004.

Z. A. Knight and K. M. Shokat. Features of selective kinase inhibitors. *Chemistry and Biology*, 12:621–637, 2005.

C. M. Krejsa, D. Horvath, S. L. Rogalski, J. E. Penzotti, B. Mao, F. Barbosa, and J. C. Migeon. Predicting ADME properties and side effects: The BioPrint approach. *Curr. Opin. Drug. Discov. Devel.*, 6:470–480, 2003.

P. Kuhn, K. Wilson, M. G. Patch, and R. C. Stevens. The genesis of high-throughput structure-based drug discovery using protein crystallography. *Curr. Op. Chem. Biol.*, 6: 704–710, 2002.

T. Lengauer, C. Lemmen, M. Rarey, and M. Zimmermann. Novel technologies for virtual screening. *Drug Discovery Today*, 9:27–34, 2004.

A. Levitzki and A. Gazit. Tyrosine kinase inhibition: an approach to drug development. *Science*, 267:1782–1788, 1995.

K. Liszewski. Drug discovery: Successful lead optimization strategies. *Genetic Engineering and Biotechonology News*, 26:14, 2006.

Y. Liu and N. S. Gray. Rational design of inhibitors that bind to inactive kinase conformations. *Nat. Chem. Biol.*, 2:358–364, 2006.

P. D. Lyne. Structure-based virtual screening: an overview. *Drug Discovery Today.*, 7: 1047–1055, 2002.

B. Ma, T. Elkayam, T. Wolfson, and R. Nussinov. Protein-protein interactions structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. USA*, 100:5772–7, 2003.

M. L. MacDonald, J. Lamerdin, S. Owens, B. H. Keon, G. K. Bilter, Z. Shang, Z. Huang, H. Yu, J. Dias, T. Minami, S. W. Michnick, and J. K. Westwick. Identifying off-targets effects and hidden phenotypes of drugs in human cells. *Nat. Chem. Biol.*, 2:329–337, 2006.

G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298:1912–1934, 2002.

S. L. McGovern, B. T. Helfand, B. Feng, and B. K. Shoichet. A specific mechanism of nonspecific inhibition. *J. Med. Chem.*, 46:4265–4272, 2003.

S. K. Mencher and L. G. Wang. Promiscuous drugs compared to selective drugs (promiscuity can be a virtue). *BMC Clin. Pharmacol.*, 5:3–9, 2005.

M. Y. Mizutani and A. Itai. Efficient method for high-throughput virtual screening based on flexible docking: Discovery of novel acetylcholinesterase inhibitors. *J. Med. Chem.*, 47:4818–4828, 2004.

M. Y. Mizutani, Y. Takamatsu, T. Ichinose, K. Nakamura, and A. Itai. Effective handling of induced-fit motion in flexible docking. *PROTEIN: Structure, Function, and Bioinformatics.*, 63:878–891, 2006.

M. R. Myers, W. He, and C. Hulme. Inhibitors of tyrosine kinases involved in inflammation and autoimmune disease. *Curr. Pharm. Des.*, 3:473–502, 1997.

David L. Nelson and Michael M. Cox. *Principles of Biochemistry*. W. H. Freeman and Company, New York, 4th edition, 2005.

M. E. M. Noble, J. A. Endicott, and L. N. Johnson. Protein kinase inhibitors: insights into drug design from structure. *Science*, 303:1800–1805, 2004.

T. Ooi, M. Oobatake, G. Nemethy, and H. A. Scheraga. Accessible surface area as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA*, 84:3086–3090, 1987.

T. I. Oprea and H. Matter. Integrating virtual screening in lead discovery. *Curr. Op. Chem. Biol.*, 8:349–358, 2004.

J. Owens. Screening: Dirty drugs' secrets uncovered. *Nat. Rev. Drug Discov.*, 5:542, 2006.

N. Pietrosemoli and A. Crespo, A.and Fernandez. Dehydration propensity of order-disorder intermediate regions in soluble proteins. *J. Proteome Res.*, 6:3519–3526, 2007.

G. M. Rishton. Failure and success in modern drug discovery: Guiding principles in the establishment of high probability of success drug discovery organizations. *Med. Chem.*, 1:519–527, 2005.

W. M. Rockey and A. H. Elcock. Rapid computational identification of the targets of protein kinase inhibitors. *J. Med. Chem.*, 48:4138–4152, 2005.

B. L. Roth, D. J. Sheffler, and W. K. Kroeze. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.*, 3:353–359, 2004.

T. Schindler, W. Bornmann, P. Pellicena, W. T. Miller, B. Clarkson, and J Kuriyan. Structural mechanism for STI-571 inhibition of Abelson tyrosine kinase. *Science*, 289:1938–1942, 2000.

S. P. Shenkin, B. Erman, and L. D. Mastrandrea. Information-theoretical entropy as a measure of sequence variability. *Proteins: Struct. Funct. Genet.*, 11:297–313, 1991.

B. B. Shoichet. Virtual screening of chemical libraries. *Nature*, 432:862–865, 2004.

S. S. Taylor and E. Radzio-Andzelm. Protein kinase inhibition: natural and synthetic variations on a theme. *Curr. Opin. Chem. Biol.*, 1:219–226., 1997.

R. Tibes, J. Trent, and R. Kurzrock. Tyrosine kinase inhibitors and the dawn of molecular cancer therapeutics. *Annu. Rev. Pharmacol. Toxicol.*, 45:357–384, 2005.

C. J. Torrance, P. E. Jackson, E. Montgomery, K. W. Kinzler, B. Vogelstein, A. Wissner, M. Nunes, P. Frost, and C. M. Discafani. Combinatorial chemoprevention of intestinal neoplasia. *Nat. Med.*, 6:1024–1028, 2000.

M. Vieth, R.E. Higgs, D.H. Robertson, M. Shapiro, E.A. Gragg, and H. Hemmerle. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta*, 1697:243–257, 2004.

P. J. Whittle and T. L. Blundell. Protein structure-based drug design. *Annu. Rev. Biophys. Biomol. Str.*, 23:349–375, 1994.

A. Wissner, E. Overbeek, M. F. Reich, M. B. Floyd, B. D. Johnson, N. Mamuya, E. C. Rosfjord, C. Discafani, R. Davis, X. Shi, S. K. Rabindran, B. C. Gruber, F. Ye, W. A. Hallett, R. Nilakantan, R. Shen, Y.-F. Wang, L. M. Greenberger, and H.-R. Tsou. Synthesis and structure-activity relationships of 6,7-disubstituted 4-anilinoquinoline-3-carbonitriles. the design of an orally active, irreversible inhibitor of the tyrosine kinase activity of the epidermal growth factor receptor (EGFR) and the human epidermal growth factor receptor-2 (HER-2). *J. Med. Chem.*, 46:49–63, 2003.